

Manfred M. Fischer
Peter Nijkamp
Editors

Handbook of Regional Science



Springer Reference

Handbook of Regional Science

Manfred M. Fischer • Peter Nijkamp
Editors

Handbook of Regional Science

With 219 Figures and 59 Tables



Springer Reference

Editors

Manfred M. Fischer

Institute for Economic Geography and GIScience
Vienna University of Economics and Business
Vienna
Austria

Peter Nijkamp

Department of Spatial Economics
Free University
Amsterdam
The Netherlands

ISBN 978-3-642-23429-3

ISBN 978-3-642-23430-9 (eBook)

ISBN 978-3-642-23431-6 (print and electronic bundle)

DOI 10.1007/978-3-642-23430-9

Springer Heidelberg New York Dordrecht London

Library of Congress Control Number: 2013936795

© Springer-Verlag Berlin Heidelberg 2013

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed. Exempted from this legal reservation are brief excerpts in connection with reviews or scholarly analysis or material supplied specifically for the purpose of being entered and executed on a computer system, for exclusive use by the purchaser of the work. Duplication of this publication or parts thereof is permitted only under the provisions of the Copyright Law of the Publisher's location, in its current version, and permission for use must always be obtained from Springer. Permissions for use may be obtained through RightsLink at the Copyright Clearance Center. Violations are liable to prosecution under the respective Copyright Law.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

While the advice and information in this book are believed to be true and accurate at the date of publication, neither the authors nor the editors nor the publisher can accept any legal responsibility for any errors or omissions that may be made. The publisher makes no warranty, express or implied, with respect to the material contained herein.

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Section Editors

Regional Housing and Labor Markets

Mark Partridge Ohio State University, Columbus, OH, USA

Alessandra Faggian AED Economics Department, Ohio State University, Columbus, OH, USA

Regional Economic Growth

Jacques Poot National Institute of Demographic and Economic Analysis, University of Waikato, Hamilton, New Zealand

Innovation and Regional Economic Development

Roberta Capello Dipartimento BEST - Building Environment Science and Technology, Milano, Italy

New Economic Geography and Evolutionary Economic Geography

Andrés Rodríguez-Pose Department of Geography and Environment, London School of Economics, London, UK

Location and Interaction

Piet Rietveld Department of Economics, Free University Amsterdam, Amsterdam, The Netherlands

Environmental and Natural Resources

Amitrajeet A. Batabyal Department of Economics, Rochester Institute of Technology, Rochester, NY, USA

Spatial Analysis and Geocomputation

Paul Longley Department of Geography, University College London, London, UK

Spatial Statistics

Peter Congdon School of Geography, Queen Mary, University of London, London, UK

Spatial Econometrics

James P. LeSage Finance & Economics, Texas State University-San Marcos, San Marcos, TX, USA

Contents

Volume 1

Section I Regional Housing and Labor Markets	1
1 Migration and Labor Market Opportunities	3
Michael J. Greenwood	
2 Spatial Equilibrium in Labor Markets	17
Philip E. Graves	
3 Labor Market Theory and Models	35
Stephan J. Goetz	
4 Job Search Theory	59
Alessandra Faggian	
5 Commuting, Housing, and Labor Markets	75
Jan Rouwendal	
6 Spatial Mismatch, Poverty, and Vulnerable Populations	93
Laurent Gobillon and Harris Selod	
7 Regional Employment and Unemployment	109
Francesca Mameli, Vassilis Tselios, and Andrés Rodríguez-Pose	
8 Real Estate, and Housing Markets	125
Dionysia Lambiri and Antonios Rovolis	
9 Housing Choice, Residential Mobility, and Hedonic Approaches	147
David M. Brasington	
Section II Regional Economic Growth	167
10 Neoclassical Regional Growth Models	169
Maria Abreu	
11 Endogenous Growth Theory and Regional Extensions	193
Zoltan Acs and Mark Sanders	

12 Incorporating Space in the Theory of Endogenous Growth: Contributions from the New Economic Geography	213
Steven Bond-Smith and Philip McCann	
13 Computable Models of Static and Dynamic Spatial Oligopoly	237
Amir H. Meimand and Terry L. Friesz	
14 Demand-Driven Theories and Models of Regional Growth	259
William Cochrane and Jacques Poot	
15 The Measurement of Regional Growth and Wellbeing	277
Philip S. Morrison	
16 Regional Growth and Convergence Empirics	291
Julie Le Gallo and Bernard Fingleton	
17 The Rise of Skills: Human Capital, the Creative Class, and Regional Development	317
Charlotta Mellander and Richard Florida	
18 Infrastructure and Regional Economic Growth	331
Arthur Grimes	
19 Spatial Policy for Growth and Equity	353
Sandy Dall'erba and Irving Llamosas-Rosas	
Section III Innovation and Regional Economic Development	373
20 The Geography of Innovation	375
Edward J. Malecki	
21 Generation and Diffusion of Innovation	391
Börje Johansson	
22 Knowledge Flows, Knowledge Externalities, and Regional Economic Development	413
Charlie Karlsson and Urban Gråsjö	
23 Clusters, Local Districts, and Innovative Milieux	439
Michaela Tripl and Edward M. Bergman	
24 Systems of Innovation and the Learning Region	457
Philip Cooke	
25 Cities, Knowledge, and Innovation	475
Frank G. van Oort and Jan G. Lambooy	
26 Networks in the Innovation Process	489
Emmanouil Tranos	

Volume 2

Section IV New Economic Geography and Evolutionary Economic Geography	505
27 Classical Contributions: Von Thünen, Weber, Christaller, Lösch	507 Roberta Capello
28 Schools of Thought on Economic Geography, Institutions, and Development	527 Philip McCann
29 New Economic Geography: Past and Future	539 Carl Gaigné and Jacques-François Thisse
30 New Economic Geography: Endogenizing Location in an International Trade Model	569 Steven Brakman, Harry Garretsen, and Charles van Marrewijk
31 Evolutionary Economic Geography and Relational Geography	591 Harald Bathelt and Peng-Fei Li
32 Path Dependence and the Spatial Economy: A Key Concept in Retrospect and Prospect	609 Ron Martin
33 Agglomeration and Jobs	631 Gilles Duranton
34 Changes in Economic Geography Theory and the Dynamics of Technological Change	649 Riccardo Crescenzi
35 Geographical Economics and Policy	667 Henry G. Overman
Section V Location and Interaction	683
36 Travel Behavior and Travel Demand	685 Kenneth Button
37 Activity-Based Analysis	705 Harvey J. Miller
38 Social Network Analysis	725 Nigel Waters
39 Land-Use Transport Interaction Models	741 Michael Wegener

40	Network Equilibrium Models for Urban Transport	759
	David Boyce	
41	Supply Chains and Transportation Networks	787
	Anna Nagurney	
42	Complexity and Spatial Networks	811
	Aura Reggiani	
43	Market Areas and Competing Firms: History in Perspective	833
	Folke Snickars	
44	Factor Mobility and Migration Models	851
	Johannes Bröcker	
45	Interregional Input–Output Models	875
	Jan Oosterhaven and Geoffrey J. D. Hewings	
46	Interregional Trade Models	903
	Geoffrey J. D. Hewings and Jan Oosterhaven	
Section VI Environmental and Natural Resources		927
47	Dynamic and Stochastic Analysis of Environmental and Natural Resources	929
	Yacov Tsur and Amos Zemel	
48	Game Theoretic Modeling in Environmental and Resource Economics	951
	Hassan Benchekroun and Ngo Van Long	
49	Economic Valuation: Concepts and Empirical Methods	973
	John B. Loomis	
50	The Hedonic Method for Valuing Environmental Policies and Quality	993
	Philip E. Graves	
51	Materials Balance Models	1009
	Gara Villalba Méndez and Laura Talens Peiró	
52	Spatial Environmental and Natural Resource Economics	1029
	Amy W. Ando and Kathy Baylis	
53	Climate Change and Regional Impacts	1049
	Daria A. Karetnikov and Matthias Ruth	
54	Urban and Regional Sustainability	1071
	Emily Talen	
55	Population and the Environment	1085
	Jill L. Findeis and Shadayen Pervez	

Volume 3

Section VII Spatial Analysis and Geocomputation	1105
56 The Practice of Geographic Information Science	1107
Michael F. Goodchild and Paul A. Longley	
57 Geospatial Analysis and Geocomputation: Concepts and Modeling Tools	1123
Michael de Smith	
58 Geovisualization	1137
Ross Maciejewski	
59 Scale, Aggregation, and the Modifiable Areal Unit Problem	1157
David Manley	
60 Spatiotemporal Data Mining	1173
Tao Cheng, James Haworth, Berk Anbaroglu, Garavig Tanaksaranond and Jiaqiu Wang	
61 Bayesian Spatial Analysis	1195
Chris Brunsdon	
62 Cellular Automata and Agent-Based Models	1217
Keith C. Clarke	
63 Spatial Microsimulation	1235
Alison J. Heppenstall and Dianna M. Smith	
64 Spatial Network Analysis	1253
David O'Sullivan	
Section VIII Spatial Statistics	1275
65 Spatial Data and Statistical Methods: A Chronological Overview	1277
Robert Haining	
66 Exploratory Spatial Data Analysis	1295
Jürgen Symanzik	
67 Spatial Clustering and Autocorrelation in Health Events	1311
Geoffrey Jacquez	
68 Ecological Inferences and Multilevel Studies	1335
Mariana Arcaya and S. V. Subramanian	
69 Spatial Dynamics and Space-Time Data Analysis	1365
Sergio J. Rey	

70	Spatial Sampling	1385
	Eric M. Delmelle	
71	Spatial Models Using Laplace Approximation Methods	1401
	Virgilio Gómez-Rubio, Roger S. Bivand, and Håvard Rue	
72	Bayesian Spatial Statistical Modeling	1419
	Peter Congdon	
73	Geographically Weighted Regression	1435
	David C. Wheeler	
74	Geostatistical Models and Spatial Interpolation	1461
	Peter M. Atkinson and Christopher D. Lloyd	
75	Spatial Autocorrelation and Spatial Filtering	1477
	Daniel Griffith and Yongwan Chun	
Section IX Spatial Econometrics		1509
76	Cross-Section Spatial Regression Models	1511
	Julie Le Gallo	
77	Interpreting Spatial Econometric Models	1535
	James P. LeSage and R. Kelley Pace	
78	Maximum Likelihood Estimation	1553
	R. Kelley Pace	
79	Bayesian MCMC Estimation	1571
	Jeffrey A. Mills and Olivier Parent	
80	Instrumental Variables/Method of Moments Estimation	1597
	Ingmar R. Prucha	
81	Limited and Censored Dependent Variable Models	1619
	Xiaokun (Cara) Wang	
82	Spatial Panel Models	1637
	J. Paul Elhorst	
83	Spatial Econometric OD-Flow Models	1653
	Christine Thomas-Agnan and James P. LeSage	
Author Index		1675
Subject Index		1703

Contributors

Maria Abreu University of Cambridge, Cambridge, UK

Zoltan Acs School of Public Policy, George Mason University, Fairfax, VA, USA

Berk Anbaroglu SpaceTimeLab, Department of Civil, Environmental and Geomatic Engineering, University College London, London, UK

Amy W. Ando Department of Agricultural and Consumer Economics, University of Illinois at Urbana-Champaign, Champaign, IL, USA

Mariana Arcaya Department of Society, Human Development and Health, Harvard School of Public Health, Harvard University, Boston, MA, USA

Peter M. Atkinson Geography and Environment, University of Southampton, Southampton, UK

Harald Bathelt Department of Political Science and Department of Geography & Program in Planning, University of Toronto, Toronto, ON, Canada

Kathy Baylis Department of Agricultural and Consumer Economics, University of Illinois at Urbana-Champaign, Champaign, IL, USA

Hassan Bencheikroun Department of Economics and CIREQ, McGill University, Montréal, QC, Canada

Edward M. Bergman Institute for the Environment and Regional Development, Vienna University of Economics and Business, Vienna, Austria

Roger S. Bivand Department of Economics, NHH Norwegian School of Economics, Bergen, Norway

Steven Bond-Smith Department of Economics, University of Waikato, Hamilton, New Zealand

David Boyce Department of Civil and Environmental Engineering, Northwestern University, Evanston, IL, USA

Steven Brakman Faculty of Economics and Business, University of Groningen, AV, Groningen, The Netherlands

David M. Brasington Department of Economics, University of Cincinnati, Cincinnati, OH, USA

Johannes Bröcker Institute of Regional Research, Department of Economics, University of Kiel, Kiel, Germany

Chris Brunsdon School of Environmental Sciences, University of Liverpool, Liverpool, UK

Kenneth Button School of Public Policy, George Mason University, MS-3B1, Arlington, VA, USA

Roberta Capello Department Architecture, Built Environment and Construction Engineering A.B.C., Politecnico di Milano, Milan, Italy

Tao Cheng SpaceTimeLab, Department of Civil, Environmental and Geomatic Engineering, University College London, London, UK

Yongwan Chun Geospatial Information Sciences, School of Economic, Political and Policy Sciences, University of Texas at Dallas, Richardson, TX, USA

Keith C. Clarke Department of Geography, University of California, Santa Barbara, Santa Barbara, CA, USA

William Cochrane School of Social Sciences, University of Waikato, Hamilton, New Zealand

Peter Congdon School of Geography, Queen Mary University of London, London, UK

Philip Cooke Centre for Advanced Studies, Cardiff University, Cardiff, UK

Riccardo Crescenzi London School of Economics, London, UK

Sandy Dall'erba Regional Economics And Spatial Modeling (REASM) Laboratory, University of Arizona, Tucson, AZ, USA

Eric M. Delmelle Department of Geography and Earth Sciences, University of North Carolina at Charlotte, Charlotte, NC, USA

Gilles Duranton Department of Economics, University of Toronto, Toronto, ON, Canada

J. Paul Elhorst Department of Economics, Econometrics and Finance, University of Groningen, Groningen, The Netherlands

Alessandra Faggian AED Economics, Ohio State University, Columbus, OH, USA

Jill L. Findeis Division of Applied Social Sciences, University of Missouri-Columbia, Population Research Institute, Pennsylvania State University, Columbia, MO, USA

Bernard Fingleton Department of Economics, University of Strathclyde, Glasgow, Scotland, UK

Richard Florida Rotman School of Management, University of Toronto, Toronto, ON, Canada

Terry L. Friesz Department of Industrial and Manufacturing Engineering, Pennsylvania State University, University Park, PA, USA

Carl Gaigné INRA, UMR1302 SMART, Rennes, France

Julie Le Gallo CRESE, Université de Franche-Comté, Besançon, France

Harry Garretsen Faculty of Economics and Business, University of Groningen, AV, Groningen, The Netherlands

Laurent Gobillon Institut National d'Etudes Démographiques (INED), PSE and CEPR, Paris, France

Stephan J. Goetz Northeast Regional Center for Rural Development and Department of Agricultural Economics, Sociology and Education, Pennsylvania State University, University Park, PA, USA

Virgilio Gómez-Rubio Department of Mathematics, School of Industrial Engineering-Albacete, University of Castilla-La Mancha, Albacete, Spain

Michael F. Goodchild Center for Spatial Studies and Department of Geography, University of California, Santa Barbara, CA, USA

Philip E. Graves Department of Economics, University of Colorado, Boulder, CO, USA

Michael J. Greenwood Department of Economics, University of Colorado, Boulder, CO, USA

Daniel Griffith Geospatial Information Sciences, School of Economic, Political and Policy Sciences, University of Texas at Dallas, Richardson, TX, USA

Arthur Grimes Motu Economic and Public Policy Research, Wellington, New Zealand

Urban Gråsjö Economics and Informatics, University West, Trollhättan, Sweden

Robert Haining Department of Geography, University of Cambridge, Downing Place, Cambridge, UK

James Haworth SpaceTimeLab, Department of Civil, Environmental and Geomatic Engineering, University College London, London, UK

Alison J. Heppenstall School of Geography, University of Leeds, Leeds, UK

Geoffrey J. D. Hewings Regional Economics Applications Laboratory, University of Illinois, Urbana-Champaign, IL, USA

Geoffrey Jacquez SUNY at Buffalo, Buffalo, NY, USA

BioMedware, Ann Arbor, MI, USA

Börje Johansson Department of Economics, Jönköping International Business School (JIBS), Jönköping, Sweden

Daria A. Karetnikov University of Maryland, College Park, MD, USA

Charlie Karlsson Jönköping International Business School, Jönköping University, Jönköping, Sweden

Dionysia Lambiri Geography and Environment, University of Southampton, Highfield, Southampton, UK

Jan G. Lambooy Department of Economic Geography, Faculty of Geosciences, Utrecht University, Utrecht, The Netherlands

Julie Le Gallo CRESE, Université de Franche-Comté, Besançon, France

James P. LeSage Department of Finance and Economics, Texas State University – San Marcos, San Marcos, TX, USA

Peng-Fei Li Department of Urban & Regional Economy and Institute of China Innovation, East China Normal University, Shanghai, People's Republic of China

Irving Llamosas-Rosas Department of Economics, University of Arizona, Tucson, USA

Christopher D. Lloyd School of Environmental Sciences, University of Liverpool, Liverpool, UK

Ngo Van Long Department of Economics and CIREQ, McGill University, Montréal, QC, Canada

Paul A. Longley Department of Geography, University College London, London, UK

John B. Loomis Department of Agricultural and Resource Economics, Colorado State University, Fort Collins, CO, USA

Ross Maciejewski Arizona State University, Tempe, USA

Edward J. Malecki Department of Geography, Ohio State University 1036 Derby Hall, Columbus, OH, USA

Francesca Mameli Dipartimento di Scienze Economiche e Aziendali and CRENOS, Università degli Studi di Sassari, Sassari, Italy

David Manley School of Geographical Sciences, University of Bristol, Bristol, UK

Charles van Marrewijk Utrecht University School of Economics, University of Utrecht, TC, Utrecht, The Netherlands

Ron Martin Department of Geography, University of Cambridge, Cambridge, UK

Philip McCann Department of Economic Geography, University of Groningen, Groningen, The Netherlands

Amir H. Meimand Department of Industrial and Manufacturing Engineering, Pennsylvania State University, University Park, PA, USA

Charlotta Mellander Jönköping International Business School, Jönköping, Sweden

Harvey J. Miller Department of Geography, University of Utah, Salt Lake City, UT, USA

Jeffrey A. Mills Department of Economics, University of Cincinnati, Cincinnati, OH, USA

Philip S. Morrison School of Geography, Environment and Earth Sciences, Victoria University of Wellington, Wellington, New Zealand

Anna Nagurney Department of Finance and Operations Management, Isenberg School of Management, University of Massachusetts, Amherst, MA, USA

Jan Oosterhaven Faculty of Economics and Business, University of Groningen, Groningen, The Netherlands

David O'Sullivan School of Environment, University of Auckland, Auckland, New Zealand

Henry G. Overman Spatial Economics Research Centre and Department of Geography and Environment, London School of Economics and Political Science, London, UK

R. Kelley Pace Department of Finance, E.J. Ourso College of Business Administration, Louisiana State University, Baton Rouge, LA, USA

Olivier Parent Department of Economics, University of Cincinnati, Cincinnati, OH, USA

Shadayen Pervez Division of Applied Social Sciences, University of Missouri-Columbia, Population Research Institute, Pennsylvania State University, Columbia, MO, USA

Jacques Poot National Institute of Demographic and Economic Analysis, University of Waikato, Hamilton, New Zealand

Ingmar R. Prucha Department of Economics, University of Maryland, College Park, MD, USA

Aura Reggiani Department of Economics, University of Bologna, Bologna, Italy

Sergio J. Rey GeoDa Center for Geospatial Analysis and Computation, School of Geographical Sciences and Urban Planning, Arizona State University, Tucson, AZ, USA

Andrés Rodríguez-Pose Department of Geography and Environment, London School of Economics, London, UK

Jan Rouwendal Department of Spatial Economics, VU University, Amsterdam, The Netherlands

Antonios Rovolis Department of Economic and Regional Development, Panteion University of Athens, Kallithea, Greece

Håvard Rue Department of Mathematical Sciences, Norwegian University of Science and Technology, Trondheim, Norway

Matthias Ruth Department of Civil and Environmental Engineering, School of Public Policy and Urban Affairs, Northeastern University, College Park, USA

Mark Sanders Utrecht School of Economics, Utrecht, The Netherlands

Harris Selod The World Bank, PSE-INRA and CEPR, Washington, DC, USA

Dianna M. Smith Queen Mary University, London, UK

Michael de Smith Department of Geography, University College London, London, UK

Folke Snickars Department of Urban Planning and the Environment, KTH Royal Institute of Technology, Stockholm, Sweden

S. V. Subramanian Department of Society, Human Development and Health, Harvard School of Public Health, Harvard University, Boston, MA, USA

Jürgen Symanzik Department of Mathematics and Statistics, Utah State University, Logan, UT, USA

Emily Talen Arizona State University, Tempe, AZ, USA

Laura Talens Peiró Social Innovation Centre, INSEAD, Fontainebleau, France

Garavig Tanaksaranond SpaceTimeLab, Department of Civil, Environmental and Geomatic Engineering, University College London, London, UK

Jacques-François Thisse CORE, Université catholique de Louvain, Louvain-la-Neuve, Belgium

CMSE, NRU–Higher School of Economics, Saint–Petersburg, Russia

Christine Thomas-Agnan G.R.E.M.A.Q., Toulouse School of Economics, Toulouse, France

Emmanouil Tranos Department of Spatial Economics, VU University, Amsterdam, The Netherlands

Michaela Tripl Department of Human Geography, Lund University, Lund, Sweden

Vassilis Tselios Geography and Environment, University of Southampton, Southampton, UK

Yacov Tsur Department of Agricultural Economics and Management, Hebrew University of Jerusalem, Rehovot, Israel

Frank G. van Oort Department of Economic Geography, Faculty of Geosciences, Utrecht University, Utrecht, The Netherlands

Gara Villalba Méndez Universitat Autònoma de Barcelona, Bellaterra, Spain

Jiaqiu Wang SpaceTimeLab, Department of Civil, Environmental and Geomatic Engineering, University College London, London, UK

Xiaokun (Cara) Wang Department of Civil and Environmental Engineering, Rensselaer Polytechnic Institute, Troy, NY, USA

Nigel Waters Department of Geography and Geoinformation Science, George Mason University, Fairfax, VA, USA

Michael Wegener Spiekermann & Wegener, Urban and Regional Research, Dortmund, Germany

David C. Wheeler Department of Biostatistics, Virginia Commonwealth University, Richmond, VA, USA

Amos Zemel Department of Solar Energy and Environmental Physics, Jacob Blaustein Institutes for Desert Research, Ben Gurion University of the Negev, Sede Boker Campus, Israel

Regional Science at Full Gallop: Editorial Introduction

Manfred M. Fischer and Peter Nijkamp

1 Aims and Scope

This Springer *Handbook of Regional Science* is meant to be a major reference work. It brings systematically together a varied set of major contributions to regional science that may be considered to be landmarks of advanced collective knowledge in the field. It was conceived to provide an understanding of major developments in regional science, in theory, methodology, and application. The various contributions are not purely theoretical or applied in nature, but offer a tertiary literature overview of advances in the field over the past few decades. The growth pace of regional science has been so fast that it is highly unlikely that a single scholar could have command of either the full spectrum of technical research tools, the broad multidimensional array of theoretical contributions, or the varied range of operational frameworks and studies in regional science.

Regional science has over the past half a century indeed turned into a broad multidisciplinary orientation on regional and urban issues, combining – and being a complement to – regional economics, social and economic geography, urban economics, transportation science, environmental science, political science, and planning theory. Regional science has also developed a powerful scientific toolbox that is nowadays being used in many spatial analyses. A major aim of the present handbook is to make major developments in regional science accessible to a broad set of students, researchers, practitioners, and teachers, as well as to provide a strategic source of reference for many interested scientists in the years to come.

It should be noted that regional science has – apart from a few notable exceptions – not been served very well with advanced textbooks, which makes access of regional science theory and method to advanced students and interested scholars rather difficult. This multi-volume handbook aims to provide a genuine and appealing entry to a rich and expanding scientific field, in which the interface of social sciences and space is highlighted from an analytical perspective.

2 A Short History

Regional science as a broadly recognized scientific domain has been brought to fruition only over the last 50 years. This does not imply that in previous periods there was no interest in spatial issues. On the contrary, already the grandfather of economics, Adam Smith, analyzed the relationship between location and trade, by emphasizing the importance of accessibility in spatial connectivity. And many other classical scholars, e.g., Ricardo, Malthus, Quesnay, and several others, have addressed – often implicitly – important issues of the space-economy. Of course, there is also a range of recognized predecessors of regional science, in particular, von Thünen, Weber, Palander, Predöhl, and Lösch. But the real history of regional science started with the seminal contributions of Walter Isard, who laid the foundation for a rigorous analytically oriented regional science since the mid 1950s. The framework developed by him had a theoretical foundation, a strong methodological orientation, and a strong emphasis on applied modeling of real-world phenomena and processes, seen from a multidisciplinary perspective.

It is noteworthy that Isard did not only provide original contributions to regional science in a strict sense, but also to ecological science, transportation science, and even conflict management. His approach is a perfect example of the multidisciplinary nature of regional science. This interdisciplinary character is also the key feature of the present handbook. Contributors of the various chapters originate from several disciplines which all together make up the constituents of regional science. These contributions follow the strict methodological requirements imposed in the early genesis of regional science, in which quantitative analysis and multidisciplinary approach are key.

A major recurrent theme in regional science is location and agglomeration theory. Location and agglomeration derive their importance from distance frictions, economies of scale, and proximity and connectivity, which are inherent in the spatial behavior of economic agents (households, firms, public actors). This theme forms the prominent historical perspective for regional science. And therefore, in the next sections we will concisely address this theme.

3 Location and Agglomeration Theory

Locations and agglomerations are spatial phenomena par excellence. They were historically – next to spatial interactions, e.g., through transport or trade – the most obvious subject matters of research in regional science. And over the decades, regional science has built up a strong tradition in analytical research on the determinants or drivers of location in the space-economy. Location theory does not only include industrial location decision, but also residential location and

facility location, including the spatial interactions between locations (allocation) and the spatial concentration of activities (agglomeration).

Location and agglomeration theory is concerned with the question where and why economic activity can be found. It addresses the spatial behavior of all agents, not only from a point perspective (i.e., the location), but also from a spatial pattern and geographical structure perspective (i.e., agglomerations and interdependencies). Individual location decisions were already studied more than a century ago by von Thünen and Weber, while geographical clusters and spatial interactions were already studied by regional scientists *avant la lettre* (e.g., Marshall, Palander, Christaller, Predöhl, Perroux, Myrdal, Hägerstrand, and many others). Regional science from a locational angle did not only spur innovative, theoretical, and methodological research on the space-economy but also applied policy research (e.g., on growth poles, industrial districts, etc.). This has also prompted a far-reaching research interest in regional development in a broad sense or regional economic growth in a more limited sense. Recent examples can be found in the endogenous growth theory, the New Economic Geography, or the neo-innovation theory.

In the same vein, housing markets and labor markets have become foci of regional science research, often from an urban economics perspective. In this context, land rent and mobility behavior are related to modern location and agglomeration analysis. Urban dynamics – including urban sprawl and the emergence of the “New Urban World” – has consequently also become a prominent direction in regional science research, along with transportation research and, more recently, digital infrastructure and geoscience research.

The reader will note that the present handbook does not contain a particular section devoted to location, allocation, and agglomeration. The editors have deliberately decided not to include a special section on these topics, for the simple reason that in the rich history of regional science location and agglomeration theories have increasingly become mainstream with more integrative spatial research themes, such as regional growth, regional innovation, spatial labor and housing markets, spatial modeling, and so forth. From this perspective, there is no evident or compelling need for a dedicated location/agglomeration section. There are many contributions in this handbook that address locational issues, but often embedded in a broader spatial context. Since “location is everywhere,” the editors feel that a separate section on location is no longer warranted.

4 Organization of the Handbook

The design of this handbook follows strict logical principles. There are nine major parts (sections), each of which consists of a set of systematically organized chapters. Though each author is responsible for the contents of his or her chapter,

a strict review procedure has been adopted, by both the section editors and the editors-in-chief. The handbook editors and the section editors have critically reviewed each individual contribution. This has not only ensured a strict quality control on each submitted chapter, but also a functional coherence and integration of all chapters and sections. And therefore, this handbook is more than a collection of loose chapters.

Clearly, in an interdisciplinary setting, a subdivision of a domain into sections and chapters is never watertight, but in our view the current structure of the handbook serves as a useful structuring of central themes in regional science. This opus contains nine overarching themes:

- Regional housing and labor markets
- Regional economic growth
- Innovation and regional economic development
- New economic geography and evolutionary economic geography
- Location and interaction
- Environmental and natural resources
- Spatial analysis and geocomputation
- Spatial statistics
- Spatial econometrics

These themes will now successively be discussed in a succinct way. This will be done in a rather novel way. Rather than offering summaries of each of the nine sections and of all 83 chapters of this handbook – which would be a boring and voluminous task – we will employ a so-called content cloud analysis which maps out in a visually appealing way the most prominent terms and concepts used in each individual section as well as in the handbook as a whole. A content cloud analysis is based on a systematic digital search algorithm, through which the most relevant substantive items – in terms of frequency – can be traced and identified, and next be included in a multicolor visualization, in which commonalities and frequencies of such items can be shown through color intensities and font sizes. This will be done here for all nine sections and for the entire opus, followed by a concise exposition. This way of systematizing the structure and context of this handbook offers also a key to trace cross-references through the subject index composed for this work.

5 **Regional Housing and Labor Markets**

Spatial housing and labor market research has been a focal point of attention in the long-standing history of regional science. Housing and labor markets are the cornerstones of regional science, as they related to both residential and firm locational behavior. They have been extensively treated in the history of spatial – urban and regional – research. The various contributions in this section extend the traditional focus in this field by including also migration, job search, poverty, real estate, and market-based evaluations. The content cloud associated with the nine chapters in this section is depicted in [Fig. 1](#).

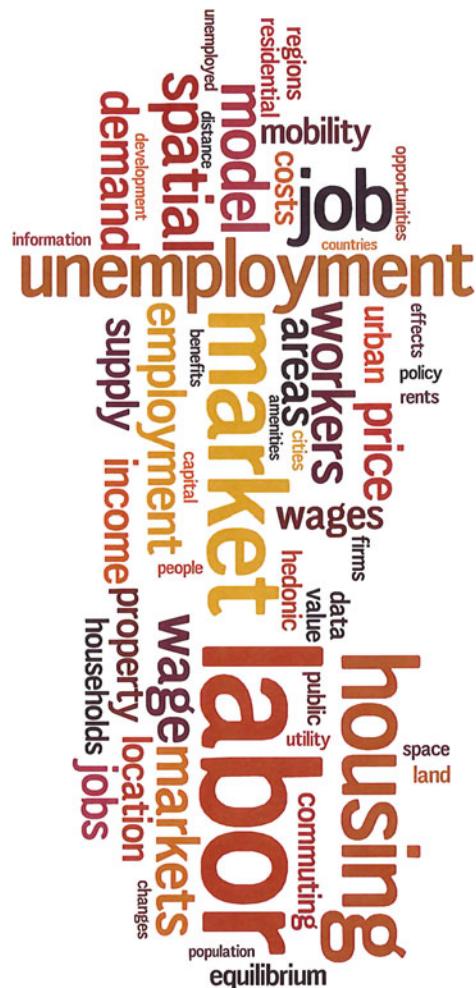


Fig. 1 Content cloud of the Section on “Regional Housing and Labor Market”

This content cloud contains evidently labor, housing markets, (un)employment, and wage(s) as prominent items. But also terms like supply and demand, property, mobility, households, rents, and locations and hedonic values are clearly present. This colorful spectrum appears to present a balanced treatment of concepts one might expect in a section on housing and labor markets.

6 Regional Economic Growth

Regional economic growth issues have inspired a wealth of research – both theoretically and politically oriented – on the drivers and implications of

imbalances in spatial development patterns. Recent advances from macroeconomic growth theory and from economic geography have once more highlighted the importance of studying the causes of spatial disparity phenomena, from both a spatial equilibrium perspective and real-world regional policy perspective. More recently, issues such as well-being, health, education and skills, and human capital have also been included in spatial equity and convergence discussions.

The content cloud related to the second section (see Fig. 2) confirms largely the aforementioned sketch of regional economic growth issues. Next to traditional items such as income, production, or equilibrium also more recently popular concepts like knowledge, well-being, innovation, and migration appear to play a substantive role in the individual chapters of this section.

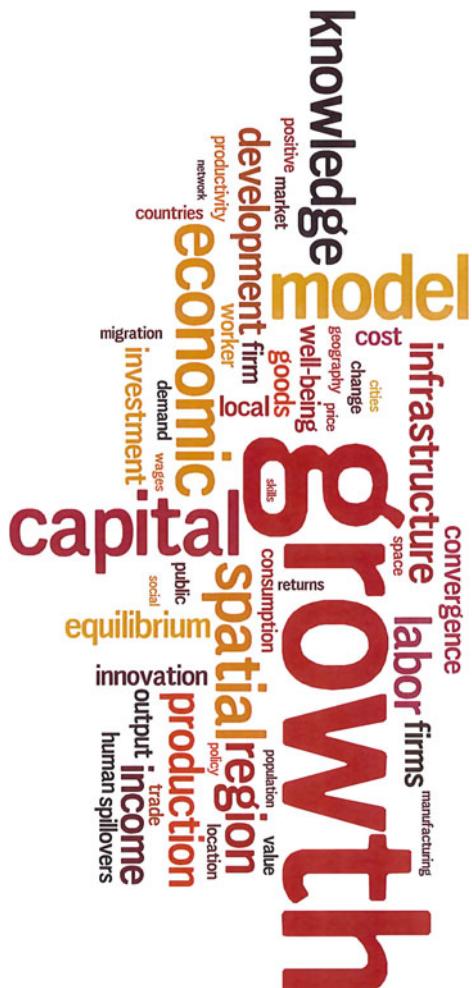


Fig. 2 Content cloud of the Section on “Regional Economic Growth”

7 Innovation and Regional Economic Development

The third section of this handbook is devoted to innovation and regional economic development. It contains various important contributions on the spatial pattern of innovations, in combination with knowledge diffusion and absorption. Important elements are in particular externalities, innovative milieux, learning regions, human capital in cities, and digital infrastructures. This section offers a wealth of systematic insights into spatial dynamics and regional development.

This is confirmed by the content cloud related to this section (see Fig. 3). Keywords which stand out here are in particular: knowledge, networks, spillovers,

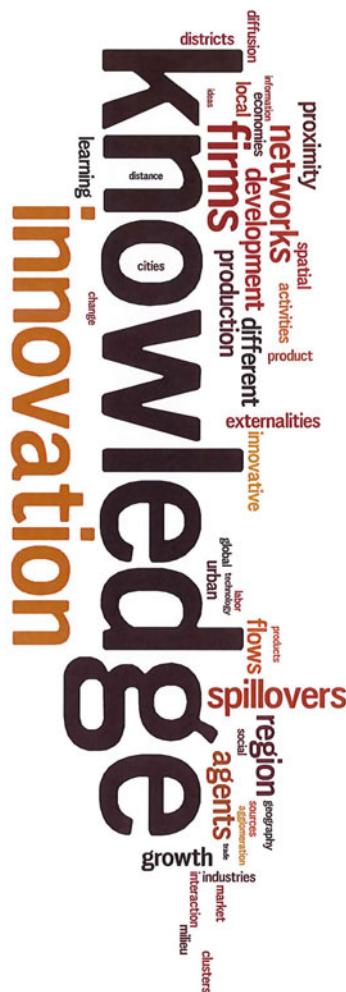


Fig. 3 Content cloud of the Section on “Innovation and Regional Economic Development”

and proximity. But also items such as interaction, diffusion, externalities, and learning are well represented. All in all, this section provides a collection of critical concepts that are key to a treatment of innovation and regional development.

8 **New Economic Geography and Evolutionary Economic Geography**

In the past decades, much attention has been devoted to new conceptualizations of spatial dynamics. The present section offers various interesting contributions on new economic geography and evolutionary economic geography. Rather than discussing whether we have here old wines in new bottles, the various chapters offer a systematic positioning of these issues in the regional science history and, more broadly, in modern economic growth theory. In this vein, also due attention is paid to institutional frameworks, to endogenous location and trade, to evolutionary perspectives and path dependencies, as well as to agglomeration externalities and the role of technological change.

The wealth of information contained in the front section of this handbook is substantiated by the content cloud mapped out in Fig. 4. Items like markets, cities, agents, location, competition, and agglomeration dominate the scene in this content cloud, followed by important terms such as innovation, knowledge, equilibrium, proximity, and clusters. This information mapping clearly confirms the solidity of the topical choice on the theme of new economic geography and evolutionary economic geography.

9 **Location and Interaction**

Spatial interdependencies have always been at the heart of regional science research. Such interdependencies are clearly related in transport flows and mobility patterns, but go also much further. The present section on location and interaction does not only offer an account of travel patterns, transportation analysis, and network models, but also provides new insights into activity-based analysis, social network configurations, and spatial land-use models. This overview is then extended toward adjacent domains, such as supply chains, complex spatial systems, market areas, trade and migration, and input-output linkages. This rather comprehensive section illustrates the rich heritage which has been gathered in the spatial modeling history in regional science.

The latter observation is confirmed by the content cloud in Fig. 5, which offers a visual mapping of key concepts in the fifth section of the handbook. Clearly, transport, mobility, networks, trade, and travel are prominently present in this cloud, but also terms like complexity, social, public, time, and change show up. This indicates that this section treats a variety of relevant concepts that may be seen as essential for a section on location and interaction.



Fig. 4 Content cloud of the Section on “New Economic Geography and Evolutionary Economic Geography”

10 Environmental and Natural Resources

Environmental and natural resources have been seen as major drivers of regional development since the early history of regional science. But its importance has not changed over the past decades, partly due to the awareness of the necessity of these resources for human survival, partly due to new emerging issues such as ecological sustainability, spatial resilience, or climate change. The present section on environmental and natural resources aims to provide up-to-date insights into the

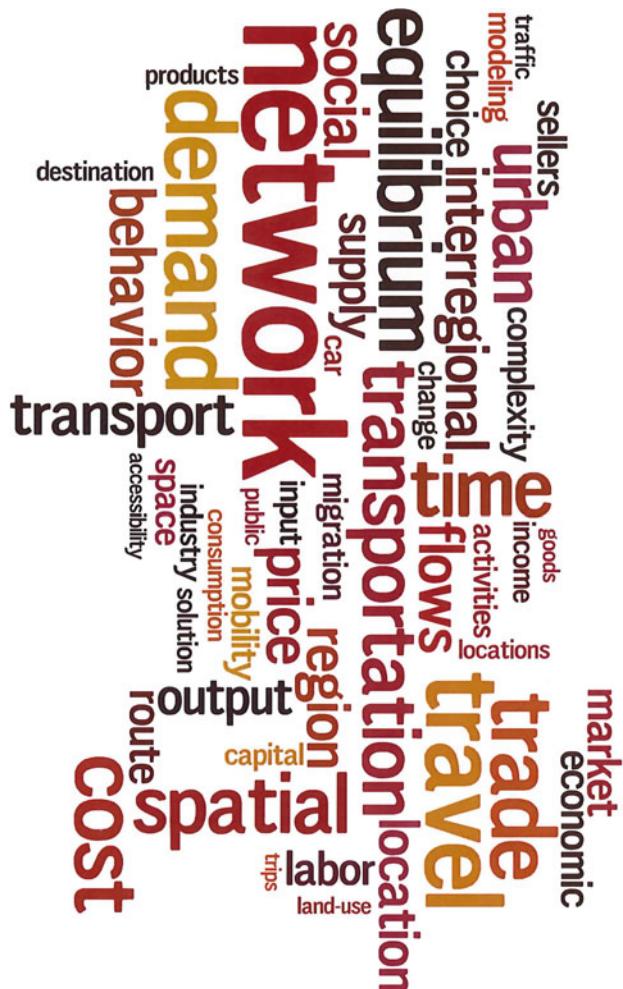


Fig. 5 Content cloud of the Section on “Location and Interaction”

development of analytical tools and conceptual frameworks in this field. Important methodological advances can be found *inter alia* in: stochastic analyses, game-theoretic frameworks, new economic valuation methods, hedonic valuation techniques, and materials balance models. New angles can also be found by an explicit consideration of spatial patterns of ecological resources, climate change, urban sustainability, and population dynamics.

The content cloud mapped out in Fig. 6 depicts most of the key concepts in this section on environmental and natural resources. Prominent concepts that show up in this figure are – apart from environmental – population, resources, emissions, growth, value, space, and location. But also related items such as conservation,



Fig. 6 Content cloud of the Section on “Environmental and Natural Resources”

uncertainty, social, and public are present. This cloud shows that the position of environmental and natural resources is well anchored in regional science.

11 Spatial Analysis and Geocomputation

The section on spatial analysis and geocomputation offers an overview of achievements in a rapidly evolving research field in regional science. It provides relevant contributions to geographic information science, geovisualization, geospatial



Fig. 7 Content cloud of the Section on “Spatial Analysis and Geocomputation”

analysis, and geocomputation. This collection of advances is complemented with a useful state-of-the-art overview on the modifiable area unit problem (MAUP), spatiotemporal data mining, and Bayesian spatial analysis. To complete this interesting overview, several chapters have also been added to deal with cellular automata, agent-based models, spatial microsimulation, and spatial network analysis. Most of these topics have attracted major research interest in recent decades and may be seen as important contributions to the methodology of regional science.

The above observations are confirmed by the information contained in the content cloud associated with this section on spatial analysis and geocomputation (see Fig. 7). Keywords in this cloud are, apart from spatial and data: network,

distribution, space-time, attributions, graphs and maps. But also related terms play a significant role, such as: scale, visualization, location, behavior, and change. All in all, the above description suggests that this section covers a wide variety of research tools and concepts in modern regional science.

12 Spatial Statistics

Spatial statistics has become a rapidly growing research area in regional science, especially in the context of exploratory spatial data analysis. This section offers interesting horizons for new insights into spatial statistical data analysis. Spatial clustering, spatial dynamics and space-time data analysis, ecological inferences, and multilevel statistical analysis are important topics considered in this eighth section. Other advanced items that are treated in this section are *inter alia*: Bayesian statistical analysis, geographical weighted regression, geostatistical modeling and spatial filtering techniques, and geostatistical modeling and spatial interpolation. It is evident that this section contains a wealth of recent insights into the research potential of sophisticated statistical research techniques in regional science.

These findings are in agreement with the results of the content cloud presented in Fig. 8. Apart from evident terms like regressions, spatial, or random, also various other items are presented, such as sampling, multilevel, and inference of software. In addition, relevant concepts like risk, variogram, eigenvectors, or neighborhood are included. This means that the present section on spatial statistics provides an extremely useful overview of the statistical toolbox of modern regional science research.

13 Spatial Econometrics

Spatial econometrics refers to the econometric toolbox applied to and adjusted specifically for spatially interdependent phenomena and processes. It has already quite a long history and gradually evolved into an important subfield within regional science. This section on spatial econometrics offers an up-to-date overview of advances in the flourishing research domain. Next to general overviews, also various specific topics are treated here, such as maximum likelihood estimation methods, Bayesian estimation techniques, instrumental variable methods, and the like. In addition, various advanced techniques are presented as well, in particular, limited and censored dependent variable models, spatial panel models, and spatial econometric origin-destination (OD) flow models. This section forms – next to the previous sections – a balanced representation of the toolkits of quantitative regional science research.

The content cloud in Fig. 9 offers a confirmation of the above-mentioned remarks. Next to emphasis on standard terms like spatial effects and spatial models, we find also a prominent position for such concepts as explanatory, distribution,



Fig. 8 Content cloud of the Section on “Spatial Statistics”

flows, specification, and dependence. Furthermore, also various other important terms come to the fore here, in particular: distance, interaction, neighbors, chains, destinations, lags, and spillovers. It goes without saying that this research field still shows rapid dynamics, with many more advances to come in the future.

14 Synthesis of Concepts

The *Handbook of Regional Science* covers a wide variety of concepts, methods, frameworks, and research tools. The terms with the highest frequency of

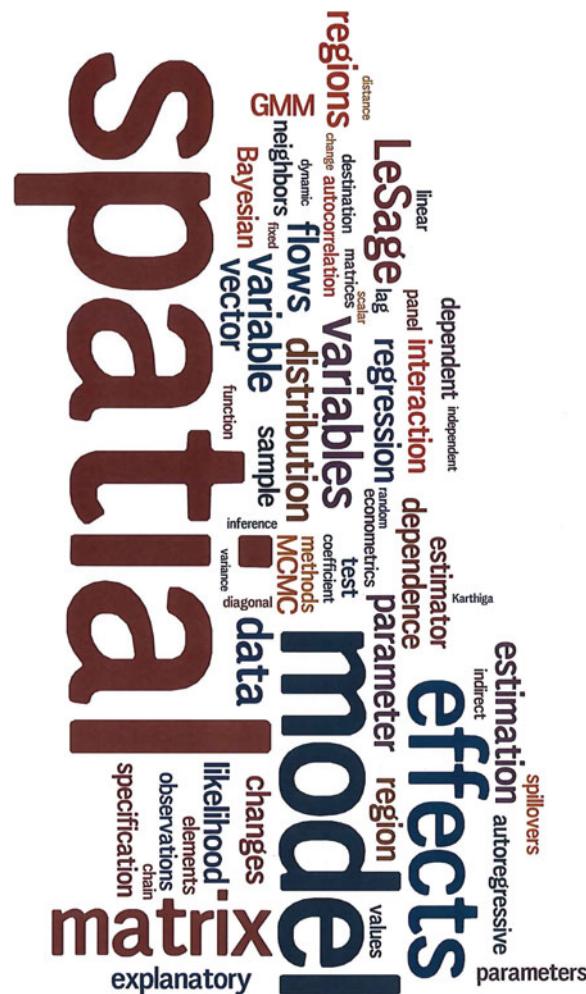


Fig. 9 Content cloud of the Section on “Spatial Econometrics”

appearance through the entire handbook are summarized in the content cloud in Fig. 10. This figure offers indeed a balanced representation of the main substance issues covered by this handbook. This is further confirmed by the content cloud in Fig. 11 which is based on a systematic screening and recording of the main concepts included in the subject index of this handbook. Both figures depict largely the same type of information and may be seen as the main ingredients of this volume.



Fig. 10 Content cloud of the *Handbook of Regional Science*

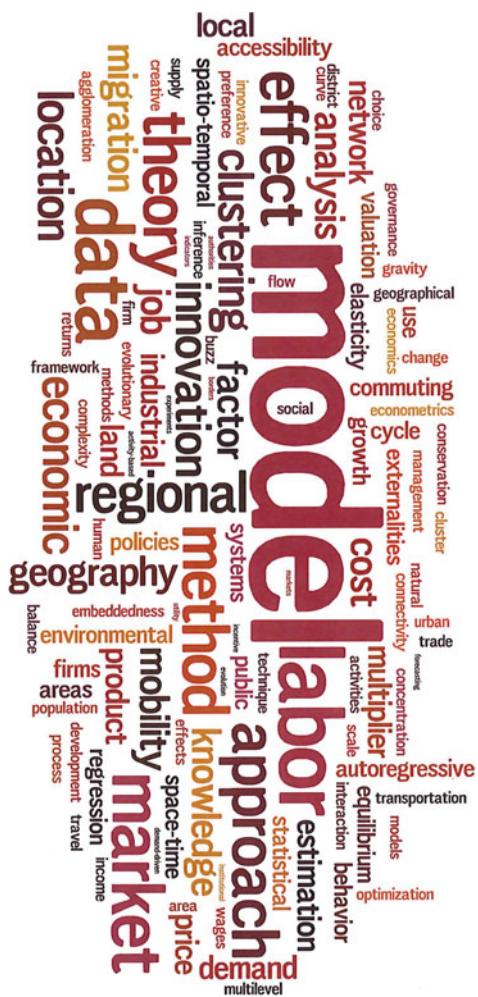


Fig. 11 Content cloud of the subject index of the *Handbook of Regional Science*

Spatial model is a model that studies the spatial distribution of economic activities and the spatial effects of economic decisions. It integrates geographical information systems (GIS), spatial statistics, and spatial econometrics to analyze the spatial patterns of economic phenomena and their causal relationships.

The spatial model is based on the assumption that economic activities are influenced by various factors such as location, capital, labor, and knowledge. These factors are often spatially correlated, meaning that they are more abundant or more expensive in certain locations than others. The spatial model attempts to capture these spatial patterns and their implications for economic growth and development.

The spatial model can be used to study a wide range of economic issues, including urbanization, regional development, labor markets, and international trade. It can also be used to predict the impact of policy changes, such as changes in taxation or regulation, on the spatial distribution of economic activities.

The spatial model is a complex and interdisciplinary field that requires a multidisciplinary approach. It involves the integration of various disciplines, such as geography, economics, and statistics, to develop a comprehensive understanding of the spatial patterns of economic phenomena.

Section I

Regional Housing and Labor Markets

Michael J. Greenwood

Contents

1.1	Introduction	4
1.2	Measuring Economic Opportunities	6
1.3	Unemployment Rates	7
1.4	Income, Earnings, and Wages	10
1.5	Employment Opportunities	12
1.6	Conclusions	14
	References	15

Abstract

This chapter traces the development of the role of economic opportunities in the study of migration. From the earliest years of internal migration as a recognized field of study, scholars in many social science disciplines believed that such opportunities were key determinants of migration. However, during the late nineteenth and early twentieth centuries, the lack of statistical measures of income and wages at subnational levels prevented empirical testing of the economic opportunity hypothesis. During this time, much rural-to-urban migration was occurring, and the presumption was that these flows were being driven by perceived urban–rural differences in economic well-being. The first formal measures used by economists in the 1930s were regional unemployment rates, and these rates proved to be significant determinants of migration during the Depression, but did not always hold up to scrutiny in later years. As aggregate income measures became increasingly available after 1960, they were incorporated in migration models, but their empirical success also was limited. Finally, the availability of microdata that

M.J. Greenwood

Department of Economics, University of Colorado, Boulder, CO, USA

e-mail: Michael.Greenwood@Colorado.EDU

reflects personal employment status and household income has allowed numerous advances in our understanding of various migration phenomena and also has helped clear up many dilemmas regarding earlier migration studies that used aggregate data.

1.1 Introduction

The earliest work on migration recognizes the importance of *economic opportunity* as a key *determinant of migration*, if not the single most important determinant. In his classic nineteenth-century article, Ravenstein (1885, p. 181) leaves little doubt that he believed employment and wage opportunities were the major “determinants” of migration: “In most instances it will be found that they did so (leave their homes) in search of work of a more remunerative or attractive kind than that afforded by the places of their birth” (parentheses are mine). Later, he wrote that “the call for labour in our centres of industry and commerce is the prime cause of these currents of migration” (p. 198). Ravenstein does, however, recognize that the motives for migration are “various.”

For many years after Ravenstein’s work, very little research focused on internal US migration, which D.S. Thomas (1938) attributes to a scholarly focus on US immigration during the period often referred to as “the age of mass migration.” However, with the imposition of binding immigration quotas in 1921 and even more restrictive quotas in 1924, followed by the Great Depression in the 1930s, immigration fell sharply, and internal migration (especially rural-to-urban, South to North, and East to West migration) claimed an important place in the study of US migration. Now economists began to focus on *internal migration* as a field of study rather than more or less exclusively on immigration. With the economists came a much more specific concern for the importance of economic opportunities as a major force underlying migration. This was a concern that they carried over from their work on international migration.

During the 1930s, in a series of articles published in *Oxford Economic Papers* that was one of the most empirically sophisticated studies of its time, Makower et al. (1938, 1939, 1940) not only anticipated the *gravity model* of migration, but they also stressed the importance of economic opportunities as measured by the unemployment rate: “Quite a close relationship was found between discrepancies in the unemployment rates and migration of labour where allowance was made for the size of the insured population and the distance over which migrants had to travel” (1938, p. 118). At about the same time, Hicks (1932, p. 76) was arguing that “differences in net economic advantages, chiefly differences in wages, are the main causes of migration.”

Ravenstein was a British geographer, whereas Hicks and Makower, Marschak, and Robinson were British economists. Understandably, their focus was primarily on Great Britain. In the United States, prominent demographer Warren Thompson was further stressing the importance of economic opportunities: “The distribution of population always has been, and still is, determined chiefly by the economic

necessities of individuals, families, or larger groups, although social usages, personal preferences, and group traditions have always interfered with the free play of the economic factors in this process” (1936, p. 250). At about the same time, economist Carter Goodrich et al. were focusing on economic opportunities during the Depression in *Migration and Economic Opportunity* (1936). Much *rural-to-urban migration* occurred during the 1930s, and the Carter group was asking questions like would the migrants be better off if they were back in the rural communities from which they had departed.

Another famous work by one of the all-time best migration researchers appeared just after the Goodrich book. This was D.S. Thomas’s (1938) *Research Memorandum on Migration Differentials*. This contribution contains surprisingly little reference to economic differences as main determinants of migration. However, Thomas clearly thought that such differences were among the top determinants: “It goes without saying that there are other important factors (among the determinants of migration) in addition to community structure, distance, and phase of the business cycle, but we regard these three as fundamental” (1938, p. 6) (parentheses mine). Her reference to “phase of the business cycle” has to refer to economic opportunities. Her focus was mainly on migration differentials, or *selective migration*, like age and sex, and it was apparently too early for her to see how economic opportunities could play a major role in the determination of who moves. For example, age selection is importantly determined by economic opportunities in the sense that migration tends to occur at early ages because to postpone moving means sacrificing the monetary returns that are discounted least.

The Carter study was conducted primarily at the University of Pennsylvania, so it is perhaps not surprising that one of the primer migration studies of the 1950s and 1960s was conducted at this University as well. Led by S. Kuznets and D.S. Thomas, the University of Pennsylvania group published *Population Redistribution and Economic Growth in the United States, 1870–1950* (1957, 1960, 1964). This research also emphasized the importance of economic opportunities: “the distribution of a country’s population at any given time may be viewed as a rough adjustment to the distribution of economic opportunities” (Kuznets and Thomas 1957, p. 2). Thus, from the very beginning of migration research as a recognized discipline for study, economic opportunities were viewed as important determinants of migration, and perhaps as the single most important set of determinants, and this view was held by scholars in several social science disciplines.

In the sense that it can be valued, either directly in the market or indirectly through imputation, almost anything may be viewed as an “economic opportunity” (Greenwood 1997). Thus, for example, *location-specific amenities*, such as desirable (or undesirable) aspects of climate, have “values” that are reflected in labor and/or land markets. However, in this chapter, my emphasis is on more traditional measures of economic opportunities. These measures include (i) *wages and incomes* and (ii) job opportunities as reflected in *employment, employment growth, unemployment rates*, and “*crowding out*.”

Although many models that concern less-developed countries are similar in their formulation to counterparts for developed countries, my focus in this chapter is

specifically on developed countries. Several very good survey articles are available on migration in less-developed countries. Especially in the context of less-developed countries, the so-called *new economics of migration* provides certain new and different perspectives on economic opportunities. In traditional approaches to migration, individuals who are presumed to be utility maximizers make the decision to migrate or not, but in this new theory, migration decisions are made by larger groups such as families. Remittances play a key role in the sense that a (*family*) member may be sent away for the express purpose of sending funds back “home.” Thus, economic opportunities are viewed in a somewhat different sense in this approach.

1.2 Measuring Economic Opportunities

In the earliest empirical studies of internal migration, economic opportunities did not play a key role because statistical measures of such opportunities simply were not available at subnational (spatial) levels. The best alternative appears to have been a focus on migration to cities, where economic opportunities were presumably seen as superior to those in the rural areas from which the migrants were coming. This orientation is clearly apparent in Ravenstein (1889) second paper and in numerous papers discussed in Thomas (1938).

Since those early days when almost no regional measures of economic well-being were available for inclusion in migration models, numerous measures have been developed and used to reflect economic opportunities. As noted above, county unemployment rates were used to study internal migration in Great Britain during the 1930s. By 1960, Easterlin (1960) had developed estimates of *US regional and state per capita income* back to 1840, as well as estimates of service income per worker at the state level. The former estimates were subsequently used in studies of historical US internal migration. In the USA, various census measures that reflected statewide mean or median income were being employed to study interstate migration. Such measures along with aggregate unemployment rates also were employed to study other geographic configurations like substate areas. Not only were such measures used to analyze primary migration, but they also were used to study secondary moves (like return and onward migration).

During the 1960s and 1970s, studies that used *aggregate place-to-place migration* measures or that studied *in- and out-migration* or *net migration* often adopted income, unemployment rates, contemporaneous employment growth (in simultaneous-equation models), and lagged employment growth (to avoid simultaneity problems). Such studies frequently used these variables defined for places of origin (to reflect forces that might push potential migrants out or encourage them to stay) and for places of destination (to reflect forces that would attract or pull migrants or, alternatively, discourage them from coming). In other instances, ratios of destination to origin variables were adopted, but these measures in the then frequently estimated double-log, *modified gravity models* constrained the coefficients on the origin and destination variables to be the same except for sign.

Before the availability of data sets like the Census Public Use Microdata Samples (PUMS), researchers were constrained to the use of aggregate measures of income and unemployment rates. Generally, they had no other option. In studies of aggregate migration, an unavoidable problem with such measures is that area averages may have little relevance to actual or even potential migrants (unless everyone is regarded as a potential migrant, which is a distinct possibility). I next turn to a discussion of some of the issues tied to the use of these aggregate measures.

1.3 Unemployment Rates

As a regional characteristic, unemployment rates presumably reflect the tightness of the regional labor market. Thus, relatively higher unemployment rates characterize regions with labor markets that should encourage out-migration and discourage in-migration. The opposite is true of regions with relatively low unemployment rates. As a personal characteristic, unemployment reflects a situation in which the individual's opportunity cost of migrating is lower and his incentive to find a job anywhere, importantly in other regions as well as in his current region of residence, is higher.

The earliest study of which I am aware that uses unemployment rates in a formal regression analysis is the Makower, Marschak, and Robinson study (1939) noted above. These economists had data from the Oxford Employment Exchange that indicated the number of persons who entered the unemployment insurance system in specific counties other than Oxford and who were residing in Oxford in 1936. Although their information included such personal characteristics as sex, age, industry of employment both before and after the move, and county of origin, they aggregated the data to the county level, presumably because at that early date, they did not know how to analyze microdata. Makower, Marschak, and Robinson defined what they called the "*relative unemployment discrepancy*" as "the ratio of the difference between the unemployment rate in the county (or Division) and the unemployment rate in the whole country, to the unemployment rate in the whole country" (1939, p. 81). Their regression results indicated that "there was a very clear correspondence between variations in the relative unemployment of the county and variations in the gains and losses by migration" (1939, p. 82). The work of these authors was important for reasons that go well beyond their use of unemployment rates in a regression analysis. They were the first researchers of whom I am aware that formally estimated a gravity model of spatial interaction, although they did not refer to their model as such. They were the first to actually estimate a distance elasticity of migration, which they called the *coefficient of spatial friction* (1938).

Focusing on approximately the same period as Makower, Marschak, and Robinson, but for the United States, Bogue et al. (1957) also provide an early regression analysis that incorporates (male) unemployment rates. They employ census data to study 1935–1940 (gross in-, gross out-, and net-) migration flows

(defined as rates) for metropolitan versus nonmetropolitan state subregions. They provide regression results for migration from both metropolitan and nonmetropolitan subregions to both metropolitan and nonmetropolitan subregions. In this analysis, unemployment rates fail to be positive and statistically significant only for total migration from nonmetropolitan areas and for migration from nonmetropolitan areas to other nonmetropolitan areas (but the signs remain positive). In their regressions for in-migration and net migration, the signs on all unemployment-rate coefficients are negative, as anticipated, and highly significant (which for these authors is 5 %).

The early multiple-regression analyses of Makower, Marschak, and Robinson and Bogue, Shryock, and Hoermann are noteworthy because they were conducted at a time before the availability of computers. They also are noteworthy because their authors obtained expected signs and statistically significant coefficients on unemployment-rate variables. For many years and for many studies after these early efforts, the results on unemployment-rate variables were not to be so uniformly “correct.” In fact, in modified gravity models, the findings associated with unemployment-rate variables were among the most consistently troublesome in terms of signs and significance levels. (“Modified (or extended) gravity models” are models in which absolute migration from i to j , or the rate of migration from i to j , is a function of the basic variables of the gravity model (distance from i to j , population of i , population of j) and additional variables, such income in i and in j , unemployment rate in i and in j , and numerous other possible variables).

Many examples are available of studies that obtain unexpected signs or insignificant coefficients on an unemployment-rate variable (Greenwood 1975a). As indicated in Greenwood (1975a, p. 403), “the failure of unemployment rates to appear to influence migration in the expected direction and/or with the expected relative magnitude has been attributed to the simultaneous-equations bias inherent in single-equation, multiple regression models. This bias is likely to be particularly marked in those studies that employ explanatory variables defined for the end of the period to analyze migration that occurred over the period, because migration is itself likely to influence end-of-period economic conditions.” In Greenwood (1975b), I examine these hypotheses with US census data on 1955–1960 and 1965–1970 metropolitan in- and out-migration, using explanatory variables defined for the beginning-of-period, end-of-period, and, alternatively, changes over the period. The models are estimated by ordinary least squares and by two-stage least squares. For the most part, no matter how the unemployment-rate variables are defined and no matter how they estimated, the coefficients tend with few exceptions to be statistically insignificant. The major exception is for metropolitan in-migration from nonmetropolitan areas, for which the sign is almost always negative and significant. When the absolute change in unemployment is included as endogenous in a simultaneous-equation model, for 1955–1960 migration, this variable tends to have the expected positive sign in the out-migration regressions and the expected negative sign in the in-migration regressions, and in both cases, the variable is significant. However, for 1965–1970 migration, the results are not so clean.

An alternative explanation for unanticipated findings on unemployment-rate variables was provided by Lansing and Mueller (1967). As stated in my 1975 article, they argue that “unemployment tends to be highest among the least mobile groups in the labor force—among persons in blue-collar occupations, among those with low skill and educational levels, and among the young. . .the unemployed tend to be workers who ordinarily would not consider migration as one of their options” (1975a, p. 403). The Lansing and Mueller hypothesis clearly relates to *personal unemployment*. However, Current Population Survey data have for many years indicated that the unemployed are more likely to migrate than the employed. Of course, such cross-tabs do not control for many other personal characteristics such as those noted by Lansing and Mueller.

The solution to much of the mystery associated with unemployment-rate variables awaited the availability of microdata. At the time my 1975 survey was written, only three of the 251 articles referenced in the paper utilized microdata. Soon after the publication of this article, the microdata revolution began in earnest. DaVanzo (1978), who used microdata from the Panel Study of Income Dynamics (PSID), provided important new insights into the unemployment-rate puzzle. Quoting from my 1985 survey, DaVanzo “shows that families whose heads are looking for work are more likely to move than families whose heads are not looking. Moreover, the unemployed are more likely to move than the employed. Higher area unemployment rates encourage the out-migration of those who are unemployed, but exert little influence on those who have a job” (1985, p. 532). This last finding has important implications for studies of (aggregate) migration that employ aggregate regional characteristics, including the regional unemployment rate. Even in the DaVanzo study, only a fraction of the unemployed actually migrate. The unemployed constitute a small fraction of the labor force and a much smaller fraction of the population, and the unemployed who actually migrate are a smaller fraction still. In aggregate studies, the numbers of individuals who actually migrate due to unemployment may simply be too small to be reflected in the empirical results.

Navratil and Doyle (1977) use *microdata* in combination with aggregate data to directly address the question of the influence of aggregation on elasticities estimated in migration models. They study 1965–1970 migration of white males, white females, black males, and black females, with 82 county groups contained within specific states serving as the observation base (with about 220,000 individuals). In one model, they use aggregate proxies for personal characteristics (like group-specific age and group-specific unemployment rate) in combination with general area characteristics (like the unemployment rate). In a second model, they replace the group-specific characteristics with personal characteristics (such as a dichotomous variable for unemployed vs. employed), and they retain the general area characteristics. The general unemployment rate is negative and significant only for white females when the first type of model is estimated and only for white males in the second (probit) type of model. Among the aggregate personal characteristics, the unemployment rates are never significant, but when they are replaced with a dummy variable reflecting personal employment status, the unemployment

variable is positive and highly significant for all four groups. Thus, individuals who were unemployed at the beginning of the migration interval (1965) were more likely to have migrated between 1965 and 1970, but in no way was such a finding possible to obtain with aggregate data alone.

1.4 Income, Earnings, and Wages

For economists, from the earliest studies, income or wage differences were considered to be the most basic of the determinants of migration. This position was strongly held with regard to both internal and international migration. Hicks's reference to the main drivers of migration noted above ("chiefly differences in wages") is a good example of the dominant position of wage or income differences in the thinking of economists. However, the empirical evidence on income and/or wages has not been uniformly in support of this hypothesis. Whereas some results strongly support the position, other evidence is more mixed. Moreover, one of the basic ideas in the neoclassical model is that migration should itself cause wage differences to narrow and migration to diminish over time, other factors held constant. However, the empirical results regarding this hypothesis also have been mixed.

Economic historians have developed a great deal of evidence that wage differences (or *wage gaps*) between the United States and various European countries were primary determinants of migration between Europe and America during much of the nineteenth and the early twentieth centuries, but the importance of wages declined as the gaps narrowed later in the nineteenth century. Thus, at least with respect to historical US immigration from Europe, empirical findings are consistent with wage gaps between the US and the European source countries providing a major impetus to migration, these gaps narrowing due importantly to the equilibrating effects of mass migration from Europe to the Americas, and in turn emigration from northern and western Europe to the Americas diminishing as the nineteenth century progressed.

Early studies of internal migration did not include wages or income, presumably because no measure was available. For example, Makower, Marschak, and Robinson include only the relative unemployment rate (along with distance and population) in their regression. Bogue, Shryock, and Hoermann employ a variable they refer to as "level of living index" (1957, p. 74). Even D.S. Thomas's famous *Migration Differentials* book contains limited reference to income or wages. In a later paper (1958) discussed below, she uses real per capita gross national product.

Contrary to historical studies of US immigration from Europe, modified gravity models of internal US migration (that became popular during the 1960s) frequently yielded unexpected signs and/or statistically insignificant coefficients on origin and destination income variables, especially on origin variables. Negative signs were typically expected on origin-income variables and positive signs on their destination counterparts, since higher income was expected to discourage out-migration and encourage in-migration. Many examples are available.

Proponents of the equilibrium hypothesis (that wage or income differences are compensating differences that reflect the values of location-specific amenities) claim to have provided an explanation for the unexpected signs and insignificant coefficients on income variables. However, even in the presence of various amenity variables, many models continued to yield unexpected findings.

One of the most understudied aspects of migration research is the *temporal relationship* between *cyclical economic activity* and migration. Although the topic has been of interest and concern for many years, good temporal data on migration prevented any in-depth analyses of the relationship. One of the Oxford studies of Makower et al. (1939) is the earliest of which I am aware that addresses the issue. Their data covered the period 1923–1937. They conclude that “the data of the Oxford study suggested that mobility increased with prosperity during the period 1933–7. . . While it suggests that mobility was reduced during the slump of 1931, it confirms the rise in mobility during the recovery. Thus mobility fluctuates in harmony with the trade cycle. It was found, further, that ‘short-distance’ movements were less sensitive to the slump than ‘long-distance’ movements” (1939, p. 94). They attribute the cyclicity of migration to out-of-pocket expenses: “in times of . . . prolonged unemployment, people find it more difficult to raise the money necessary for migrating” (1938, p. 118).

Another early study that attempted to uncover the relationship between cyclical activity and interstate migration was conducted by D.S. Thomas (1958). She and her research team at the University of Pennsylvania had developed fairly detailed state-specific net migration estimates by age and sex for each intercensal decade from 1870 to 1950. As indicated above, in her famous monograph on differentials in migration, she had noted that “phase of the business cycle” was one key to understanding migration differentials, but then she provided little or no empirical support for her hypothesis. With more detailed migration estimates, she now returned to this relationship. Her measure of economic activity was novel. She fit “six successive thirty-one year linear trends to annual data on gross national product per capita, in constant prices, beginning with the first year of each decade, cumulating the absolute deviations from each trend over each decade, and expressing their algebraic sums as percentages of corresponding cumulative trend values” (1958, p. 317). She then classified decades between 1880 and 1940 as relatively high versus relatively low in terms of economic activity. Her basic conclusion was that “young males, seeking economic betterment, (showed) a correspondingly greater *intensity* of migration during high than during low activity periods” (1958, p. 319) (parentheses mine).

Later, Greenwood et al. (1986) used annual (1957–1975) data from the One Percent Continuous Work History Sample of the US Social Security Administration to study the linkage between employment change and migration. They conclude that in an average year two additional local jobs attract about one additional employed migrant. However, like Thomas, they also find that the migrant-attractive power of an additional job behaves cyclically, rising during upswings and falling during downswings. They speculate that when the costs of migration are relatively high, such as during cycle troughs, a greater degree of *migrant self-selection* occurs,

and thus, migrant quality in terms of human capital rises. The opposite occurs during cyclical upswings. More recently, Saks and Wozniak (2011) examine long-distance migration over the business cycle. Using a variety of data sets, they too conclude that migration is procyclical, which they attribute to greater net benefits to moving during cyclical upturns. Moreover, similar to the earlier finding of Thomas, they argue that younger workers are especially procyclical in their migration behavior, presumably due to better economic conditions during cycle upswings.

In the end, we should recognize that the “*net wage*” or the “*net income differential*” is critical. Such a net value is corrected for state and local tax differences as well as differences in the values of publically provided goods and services. In the USA, state income *taxes* vary from none to significant percentages of taxable income, and many benefits also differ greatly across states and localities. Such *benefits* include differences in per student expenditures on K-12 education, as well as assistance for needy families in the form of food stamps, housing, and other services. A number of studies have addressed the importance of various types of state and/or local taxes and/or public expenditures in migration decisions, and I will not treat this literature in any detail here.

1.5 Employment Opportunities

Among those variables that reflect economic opportunities, the most consistently significant are those that proxy the availability of *jobs*. This condition tends to be true whether employment opportunities are measured as contemporaneous employment change (in simultaneous-equation models), lagged employment change, or employment rate.

Several studies of historical migration, as well as a number of those dealing with contemporary migration, assert that migrants crowd out others in the sense that the migrants encourage the out-migration of prior residents or discourage the in-migration of potential new residents. Presumably, underlying this phenomenon is competition for jobs, although cultural and other differences between immigrants and natives also could be responsible in part. The historical studies have focused on the manner in which the location of immigrants in northern US cities influenced South to North migration of the native born. During the late nineteenth and early twentieth centuries, as immigration from Europe surged, migration from South to North ebbed, and when immigration ebbed, this internal flow surged. Collins (1997) provides strong empirical support for this phenomenon. Thus, historically, broad regional growth patterns were significantly affected by immigration and by immigrant settlement patterns. Even during the Great Depression, when immigration was very low and immigrants were not a major issue, internal migrants to US cities caused the out-migration of longer-term residents. Boustan and Fishback (2010) show that during the 1930s for every 10 new migrants, 1.9 residents departed; moreover, another 2.1 individuals were unable to find a relief job, and 1.9 more moved from full-time to part-time work.

A number of studies have examined the location of the foreign born and the internal migration of the native born in the United States. The basic conclusion of

much of this work is nicely summarized in the early study by Filer (1992, p. 267): “It is clear that there is a strong relation between the arrival of immigrants in a local labor market and the mobility patterns of native workers. The higher the concentration of recent immigrants in an area, the less attractive that area appears to have been for native workers.” Filer’s focus is on Standard Metropolitan Statistical Areas and 1975–1980 net internal migration flows.

More recent studies examine various groups of internal migrants and more recent periods (such as 1985–1990 and 1995–2000), but many of these studies arrive at essentially the same conclusion. For example, Frey and Liaw (2005) use *microdata* and logit analysis to study both in and out interstate migration patterns for 1995–2000 and, after controlling for numerous personal and area characteristics, conclude that “Our results generally show no race-specific flight of whites alone from (states with large numbers of low-skilled immigrants), but rather show an accentuated out-migration of all race-ethnic groups from states with . . . high levels of foreign-born immigration” (p. 246, parentheses mine). Moreover, they find that “for every 100 new low-skilled immigrants to California there would be a net out-migration of fifty-one low-skilled domestic migrants” (p. 213). Similarly, Borjas (2006), using data from the 1960, 1970, 1980, 1990, and 2000 censuses and focusing on various skill groups, finds a powerful effect of immigrants on native internal migration amounting to two fewer natives wishing to live in a state if ten more immigrants settle there. The effect is somewhat greater at the metropolitan level, and more in line with the estimates of Frey and Liaw.

This literature is related to an old issue in migration. What is the *migrant-attractive power* of a job? Perhaps the earliest study to directly address this question is Muth’s (1971) “Migration: Chicken or Egg?” study. He found that three more jobs attracted two more employed net migrants. This was the direct effect of an additional job and does not take into account the indirect and induced effects that result from the migrant’s influence on jobs. However, this is only one possible outcome regarding the direct effect of employment on migration. Consider the following relationship: $M = f(\Delta E)$, where M refers to employment migration and ΔE refers to change in employment. The possibilities are as follows: (a) one more job attracts one more employed migrant, or $\partial M / \partial E = 1.0$; (b) another job attracts no migrants, or $\partial M / \partial E = 0$; and (c) another job attracts some fraction of a migrant (or, say, 100 jobs attract between 1 and 99 migrants), or $0 < \partial M / \partial E < 1.0$. This last case reflects Muth’s finding that, for example, 100 jobs would attract 67 employed migrants. This is the most likely case. In case b, local residents take all of the *incremental jobs*, as could be the case for less-skilled jobs such as might be available at McDonalds. On the contrary, in case a, where migrants fill all the incremental jobs, the jobs may be highly specialized, such as airplane mechanics. Little research has been done on the migrant-attractive power of different types (occupations) of jobs.

Although the most common finding is the one-to-one relationship, depending upon the specific region, other findings are evident for US regions (Greenwood and Hunt 1984). Thus, some *crowding out* appears to occur, but it is not a universal phenomenon. Even with respect to US immigrants, some studies deny the existence

of such a crowding-out relationship. Butcher and Card (1991) argue that this general conclusion is limited to New York, Los Angeles, and Miami. Based on their use of CPS data for the 1980s, they conclude that for 21 other cities, “native immigration flows during the 1980s were positively correlated with inflows of recent immigrants” (1991, p. 294). Similarly, Kritz and Gurak (2001) find little support for the hypothesis that native men migrate away from states with heavy immigrant concentrations over the 1985–1990 period.

1.6 Conclusions

Early studies acknowledged the importance of economic opportunities as a key determinant of migration, but due to lack of data reflecting such opportunities, these studies provided little empirical support for the hypothesis. A strong focus on rural-to-urban migration was evident in very early migration research, and the general assumption presumably was that economic opportunities in cities were sufficiently better than in rural areas to generate a strong flow of migrants from one type of area to the other. D.S. Thomas’s (1938) effort in the 1930s to summarize and synthesize the migration literature is a good example of this lack of data. Although she mentions the importance of the business cycle in determining the volume and age composition of migration, she never specifically addresses what she might have referred to, but did not, as “income differentials” and none of the many studies she cites introduce a measure of income or wages in an analysis of migration.

The first study that formally introduced a measure of economic opportunity was conducted in the late 1930s and used an unemployment measure to study intercounty migration in Britain. Beginning in the 1950s, as publicly provided aggregate income measures (such as median and mean income for states and more narrowly defined local areas) became more commonly available, such measures were introduced into migration models, which would have been judged severely lacking in their absence. Studies from this period are reviewed in Greenwood (1975a, 1985). Although it is fair to argue that these studies did not always find strong support for the hypothesis that income or wage differences were among the most important determinants of migration, on balance such measures did hold up reasonably well to empirical scrutiny.

Several recent studies have identified “crowding out” as a reason for internal US migration. This phenomenon relates to one group absorbing local jobs and thereby displacing others from their positions, and thus causing the displaced individuals to migrate from the area. Most frequently, new immigrants are seen as crowding-out natives, but the same relationship has been observed in the past as immigrants discouraged native out-migration from the South to the North, and during the Depression, rural-to-urban migrants crowded urban dwellers from their cities of residence.

As microdata became more widely available during the 1980s and beyond, later studies incorporated individual and household income data to allow the further and deeper study of the importance of income in the analysis of migration. For example, now spousal incomes are available that allow the study of family migration.

Thus, the empirical implementation of measures of economic opportunities in migration models was highly dependent upon the development of various measures of economic well-being, at first from the census and later from various special surveys. Without question, economic opportunities are now central to virtually any model reflecting human migration. Moreover, among economic opportunity measures, the availability of jobs stands out as the single most consistent variable to which migrants respond.

References

- Bogue DJ, Shryock HS Jr, Hoermann SA (1957) A regression analysis of factors explaining the size and composition of migration streams. In: Subregional migration in the United States, 1935–40, vol I, Streams of migration between subregions. Scripps foundation, Oxford, pp 64–76
- Borjas GJ (2006) Native internal migration and the labor market impact of immigration. *J Hum Res* 41:221–258
- Boustan LP, Fishback PV (2010) The effects of internal migration on local labor markets: American cities during the great depression. *J Lab Econ* 28:719–746
- Butcher KF, Card D (1991) Immigration and wages: evidence from the 1980s. *Am Econ Rev* 81:292–296
- Collins W (1997) When the tide turned: immigration and the delay of the great black migration. *J Econ Hist* 57:607–632
- DaVanzo J (1978) Does unemployment affect migration? – evidence from micro data. *Rev Econ Stat* 60:504–514
- Easterlin RA (1960) Interregional differences in per capita income, population, and total income. In: Conference on research in income and wealth, trends in the American economy in the nineteenth century, studies in income and wealth XXIV. Princeton University Press, Princeton, pp 1840–1950
- Filer RK (1992) The effect of immigrant arrivals on migratory patterns of native workers. In: Borjas GJ, Freeman R, Borjas GJ, Freeman R (eds) *Immigration and the workforce*. University of Chicago Press, Chicago, pp 245–270
- Frey WH, Liaw KL (2005) Migration within the United States: role of race-ethnicity. *Brookings-Wharton Papers on Urban Affairs* 6:207–262
- Goodrich C et al (1936) *Migration and economic opportunity*. University of Pennsylvania Press, Philadelphia
- Greenwood MJ (1975a) Research on internal migration in the United States: a survey. *J Econ Lit* 13:397–433
- Greenwood MJ (1975b) Simultaneity bias in migration models: an empirical examination. *Demography* 12:519–536
- Greenwood MJ (1985) Human migration: theory, models, and empirical studies. *J Reg Sci* 25:521–544
- Greenwood MJ (1997) Internal migration in developed countries. In: Rosenzweig MR, Stark O (eds) *Handbook of population and family economics*, vol 1B. Elsevier, Amsterdam, pp 647–720
- Greenwood MJ, Hunt GL (1984) Migration and interregional employment redistribution in the United States. *Am Econ Rev* 74:957–969
- Greenwood MJ, Hunt GL, McDowell JM (1986) Migration and employment change: empirical evidence on the spatial and temporal dimensions of the linkage. *J Reg Sci* 26:223–234
- Hicks JR (1932) *The theory of wages*. Macmillan, London
- Kritz MM, Gurak DT (2001) The impact of immigration on the internal migration of natives and immigrants. *Demography* 38:133–145

- Kuznets S, Thomas DS (eds) (1957, 1960, 1964) *Population redistribution and economic growth: United States 1870–1950*. American Philosophical Association, Philadelphia
- Lansing JB, Mueller E (1967) The geographic mobility of labor. Survey Research Center, Institute for Social Research, University of Michigan, Ann Arbor
- Makower H, Marschak J, Robinson HW (1938) Studies in mobility of labour: a tentative statistical measure. *Oxf Econ Pap* 1:83–123
- Makower H, Marschak J, Robinson HW (1939) Studies in mobility of labour: analysis for Great Britain, part I. *Oxf Econ Pap* 2:70–93
- Makower H, Marschak J, Robinson HW (1940) Studies in mobility of labour: analysis for Great Britain, part II. *Oxf Econ Pap* 4:39–62
- Muth RF (1971) Migration: chicken or egg? *South Econ J* 37:295–306
- Navratil FJ, Doyle JJ (1977) The socioeconomic determinants of migration and the level of aggregation. *South Econ J* 43:1547–1559
- Ravenstein EG (1885) The laws of migration. *J R Stat Soc Part 1* 48:167–227
- Ravenstein EG (1889) The laws of migration. *J R Stat Soc Part 2* 52:241–301
- Saks RE, Wozniak A (2011) Labor reallocation over the business cycle: new evidence from internal migration. *J Lab Econ* 29:697–732
- Thomas DS (1938) Research memorandum on migration differentials. Social Science Research Council, New York
- Thomas DS (1958) Age and economic differentials in interstate migration. *Popul Index* 24:313–325
- Thompson WS (1936) The distribution of population. *Ann Am Acad Pol Soc Sci* 188:250–259

Philip E. Graves

Contents

2.1	Introduction	18
2.2	The Traditional Labor Economics View of Spatial Labor Markets	19
2.3	An Urban/Regional View of Spatial Labor Markets	20
2.4	Spatial Labor Market Equilibrium in the Urban/Regional View with Suggestions for Future Research	23
2.5	Conclusions	32
	References	33

Abstract

Over long periods of human history, labor market equilibrium involved movements from low-wage areas to high-wage areas, a form of arbitrage under the implicit view that wage differentials corresponded to utility differentials. This “labor economics” view is likely to be viable as long as movement and information costs are high, and under this view, the movements would be expected to cause wage convergence over space. In recent decades, beginning as early as the 1960s in the United States, both the out-of-pocket and psychological costs of movement have plummeted with advances in transportation and communication technology and innovation. In addition, these same advances have enabled individual households and firms to have vastly improved information about potential benefits of locating in a host of potential locations. These observations, along with recent failures to observe convergence in wage rates, suggest that an alternative view – assuming a utility equilibrium over space – might better predict and explain the labor market equilibrium. This “urban/regional economics” view takes wages and rents as being compensatory for varying levels of

P.E. Graves

Department of Economics, University of Colorado, Boulder, CO, USA
e-mail: philip.graves@colorado.edu; philipegraves@gmail.com

household and firm amenities. In this view, whether the spatial equilibrium in labor markets involves convergence or divergence becomes quite a complicated issue. This chapter explores a number of the complexities, hinting at a broad range of potentially fruitful future research.

2.1 Introduction

People have been moving from location to location for millennia. Such moves were prompted by a myriad of motivations, ranging from famine, war, and religious persecution to the present concern with spatial equilibrium in labor markets. With primitive transportation technology, moves – particularly moves covering long distances – were very costly. In some cases, such as slavery, the moves were involuntary, while in other cases the moves, while voluntary, involved indentured or redemptioner servitude as a means of paying the high costs of passage. Chiswick and Hatton (2002) review the determinants and consequences of intercontinental migration over four centuries, with an emphasis on the colonial and postcolonial period. Rosenbloom (2002) provides an excellent review of the development of labor markets in the USA in a primarily cliometric framework. Both of the preceding provide a more general background for this chapter, but the central purpose here is to explore the role of migration (i) as a response to labor market variables and (ii) as a cause of change in labor market variables.

We shall see that the labor economics literature's traditional view of spatial equilibrium in labor markets has clear implications about the role of migration along these two dimensions. As emphasized in the early work of Borts and Stein (1964), (urban) regions with high capital to labor ratios would be expected to have higher wages than (rural) regions with low capital to labor ratios, leading to the expectation that labor flows would be from the latter to the former. These theoretical observations suggest that migration occurs as a response to arbitrageable variations in wage rates, a form of human capital investment. Examining the migration choice in detail, Sjaastad (1962) argues that migration would be expected to occur when the present value of the benefits of a move exceeds present value of the costs of that move. As we shall see in Sect. 2.2, under this approach the benefits of a move have been taken to be the higher wages obtainable in a potential destination.

That is, higher wages in a location are presumed to correspond to higher utility levels in that location. The long-run spatial equilibrium in this view of migration would be one of *convergence* in wage levels over space. Thus, (i) migrants respond to higher wages in a location with in-migration to that location, and (ii) the resulting in-migration reduces the divergence in wage levels over space.

A newer approach derives its insights much more explicitly from the urban/regional economics literature, rather than the labor economics literature briefly described above. The early theory of this more general approach is developed in Graves and Linneman (1979), with an early empirical quasi-dynamic application of the comparative-static theory in Graves (1979). The theory did not receive widespread attention until the now-classic contribution by Roback (1982) and the

supporting empirical work by Blomquist, Berger, and Hoehn (1988). In this approach, wages are only one of many things, notably rents and natural and man-made amenity levels, which vary over space. Roback explores the implications of assuming that spatial equilibrium is in existence, examining the nature of that equilibrium, an equilibrium in which utility levels and firm profitability are presumed equal everywhere. In this view, as we shall see in Sect. 2.3 below, wages and rents are expected to offset variations in amenities; hence, there is no longer an expectation of wage convergence, although it will become apparent that rent changes over time complicate an understanding of this prediction. That is, if rents tend, in the long run, to capture the full value of amenities (both household and firm amenities), then one would expect wage convergence to occur. The motivations for convergence will be seen, however, to be markedly different in the two approaches in terms of understanding the nature of the spatial labor market.

2.2 The Traditional Labor Economics View of Spatial Labor Markets

To tersely characterize the labor economics approach to migration, let utility be a function of goods consumption, as is implicit in the most basic texts in economics:

$$U = U(X) \quad (2.1)$$

where X is a vector of goods consumed, with U increasing in each element of X (any “bad” is redefined as a “good,” e.g., pollution becomes pollution abatement). [Equation \(2.1\)](#) is maximized subject to the usual budget constraint:

$$\begin{aligned} \text{Max } U &= U(X) \\ \text{s.t. } Y &= PX \end{aligned} \quad (2.2)$$

where P is the vector of prices corresponding to each element of the goods vector and Y is income of the decision-maker. This utility function, *increasing* in its arguments, can be written as an “indirect” utility function, V :

$$\text{Max } V = V(P, Y) \quad (2.3)$$

where V is *decreasing* in P and *increasing* in Y . Taking labor supply to be fixed, for simplicity (and to make some comparisons in the next section), the i th migrant’s location decision among the j locations becomes:

$$\text{Max } V_{ij} = V(P_{ij}, Y_{ij}) - C_{ij} \quad (2.4)$$

where migrant i faces prices, specific to his/her consumption bundle, which vary over space, and faces incomes, specific to his/her human capital, which also vary over space. The C_{ij} term measures the monetary and psychological costs of moving to location j , where this cost, too, is specific to migrant i (i.e., age, wage rate, distance to location j , affinity to friends or relatives, and the like can vary among migrants causing the costs of movement to vary).

In this simple model, potential migrants would not move if the cost of moving from their initial location, j_0 , exceeded the gain from moving to any alternative location $j \neq 0$. If a move occurs, it would be to the location with the highest net benefit of movement, j^* . In an even more simplified world with all goods being tradable at negligible transportation costs, prices would not vary and potential migrants would move to the locations with highest net income ($Y_j - C_j$).

The preceding simple model predicts that movement would generally occur from low-wage to high-wage areas, and aggregating to the local labor market level, the movement would tend to lead toward wage convergence. Low-wage locations would lose labor supply as households moved away, causing wages to rise; high-wage locations would experience increases in labor supply, causing wages to fall, a generalization of Borts and Stein's characterization of rural to urban labor flows. A more detailed discussion of how well this model predicts relative to its alternative in Sect. 2.3 comes later; for present purposes, it is sufficient to note that non-convergence, or implausibly slow convergence, is frequently observed. To the extent that there has been convergence, it has slowed or reversed beginning in the 1980s. Moreover, in areas with high levels of *net* immigration, there are also high levels of out-migration which would seem difficult to reconcile with the simple model of this section.

2.3 An Urban/Regional View of Spatial Labor Markets

In light of the observation that convergence appears to be occurring either not at all or at a pace which is implausibly slow, might there be an alternative way of looking at labor markets that is consistent with this observation? Roback (1982) provides an alternative view of labor market functioning that relies on assumed equilibrium in utility levels over space. The underlying notion is quite simple – just as there is no such thing as a “fast lane” on a freeway during rush hour – there is no such thing as a “nice place” to live vis-à-vis other places. Movement would be expected to occur to approximately equate speed in each lane, in the first case, and in the latter case, movement would be expected to occur to make all locations approximately equally desirable. Movement to the “nice” place should continue to occur until either high housing prices or low wages make that place no nicer than elsewhere. The details of the argument are, however, not quite this simple (see Taylor 2008 for the formal equations for this model corresponding to those for the disequilibrium model of Sect. 2.2). Our treatment here will be more intuitive, driven by words.

Fig. 2.1 Wage (W) and rent (R) expectations under various amenity assumptions

	Good for Firms	Neutral for Firms	Bad for Firms
Good for Households	$R > R_o$ $W?$	$R > R_o$ $W < W_o$	$R?$ $W < W_o$
Neutral for Households	$R > R_o$ $W > W_o$	(Base Case) $R = R_o$ $W = W_o$	$R < R_o$ $W < W_o$
Bad for Households	$R?$ $W > W_o$	$R < R_o$ $W > W_o$	$R < R_o$ $W?$

Locations offer amenities, which may be either natural or man made, that affect utility in the case of households (e.g., desirable weather, scenic views, restaurant diversity and quality) and affect production functions in the case of firms (e.g., deep-water ports, access to mine mouth, right to work laws). Considering these cases separately, if a location is unusually desirable to households, they would be expected to enter driving up housing demand and increasing the supply of labor. Hence, in equilibrium, a nice location should have higher housing costs (property values or rents) and lower wages. Conversely, an undesirable location would be expected, in equilibrium, to have some mix of lower housing costs and higher wages.

Turning to firm amenities, desirable locations would lead to firm in-migration, and that movement would be expected to increase demand for land (directly for industrial sites and indirectly for new employee housing) and increase the demand for labor. Hence, a desirable firm location, relative to others, would be expected to have – other things equal – higher housing prices (property values and rents) and higher wages. Similarly, locations that are undesirable to firms would be characterized by lower housing prices and wages.

Temporarily assuming homogeneity for simplicity, where all households possess the same utility functions and all firms possess the same production functions, all locations will have amenity bundles that would be, on net, reliably characterized as in Fig. 2.1.

Note that in the table, the “base-case” location (neither good nor bad for either firms or households) is seen in the center cell, where the rent is R_o and the wage is W_o . The wage and rent combination in this cell may be thought of as representing an “average” location. All other cases represent locations with amenity bundles that are either good or bad for one or both of households and firms relative to this average location. The cases merit consideration in more detail.

If a location is both good for firms and good for households, this location will clearly become quite large with much higher than average rents since both firms and households will want to move in (e.g., San Francisco Bay area). The impact on

wages will, in general, be ambiguous since the influx of households would lower wages, while the influx of firms would raise wages. This ambiguity is reflected in the $W_?$ symbol of cell 1, [Fig. 2.1](#) – without more information on the relative importance of the amenity to firms versus households, it is not possible to know what wages will be like compared to ordinary (base-case) locations. Conversely, in the bottom right cell, a location that is undesirable for both firms and households will have unambiguously lower rents but wages that may be higher or lower than W_0 , depending on relative undesirability.

A location that is bad for households and good for firms (lower left cell in [Fig. 2.1](#)) will have unambiguously higher wages in equilibrium (smaller supply of labor and larger demand for labor reinforce each other in raising wages). The impact on rents, $R_?$, is now ambiguous, since without further information about the relative magnitude of the (dis)amenity, we do not know whether the location will be larger or smaller than average. Conversely, in the upper right cell, if a location is bad for firms and good for households, it will have unambiguously lower wages, with an ambiguous effect on rents, because the relative importance of the (dis)amenity is not known without further information from empirical investigations.

A confusion that persists in the general population, and to a lesser extent among economists, is the role of “cost of living” in labor market equilibrium. In the case of a location that is very desirable to households and neutral to firms, the higher rents are not a higher “cost” of living but rather a higher “benefit” of living, just one that we, perhaps unfortunately, have to pay for. A nice location vis-à-vis an undesirable location is exactly analogous to comparing a new BMW to a 1980s GMC K-car – you pay more for the former, but you get more. Note however that when an amenity is neutral for households but desirable for firms, the higher rents *do* reflect a higher “cost of living.” However, in equilibrium, that higher cost of living must be compensated for by higher wages – households can be no worse off financially in such locations in equilibrium, because such amenity-neutral locations would otherwise be less desirable than alternatives.

From this point on, it is assumed that disequilibrium view of labor economics in [Sect. 2.2](#) is no longer an appropriate way to view spatial labor market equilibrium (see Partridge [2010](#) for a recent gathering of corroborative evidence). As discussed at the outset, it is likely that arbitrageable variation in real utility was the dominant cause of labor market flows when costs of movement were high and when information about the benefits of movement was costly, if available at all, as with transcontinental moves for many centuries. Costs of movement have fallen dramatically, especially relative to income growth in the United States (e.g., Interstate Highway System, more widely available and reliable automobiles, falling airfares, and long-distance phone rates to maintain psychological and other connections). Rapid advances in information technology (e.g., television beginning in the 1950s, ubiquitous by the 1960s, internet) have resulted in Americans in all locations knowing a great deal about the general nature of many if not most potential destinations. Additionally, as made clear by Roback, it is now apparent that the labor markets and the land markets cannot be considered separately, since an

amenity will generally be capitalized into both markets, something completely ignored in the labor market arbitrage model.

Two early empirical findings helped initiate the shift from the labor market disequilibrium approach. Graves (1979), using data from the 1960s, found net migration was occurring to locations with *lower* incomes, not to locations with higher incomes. Only when climate amenities were introduced into the regression did the income variable take on its “proper” sign; clearly if all amenities could be held constant in an estimating equation, more income would have to be preferred to less. Moreover, in the context of the labor market disequilibrium approach, higher rents would represent higher “costs of living,” hence would, *ceteris paribus*, be expected to lead to lower real incomes – one should expect movement, holding nominal income constant, away from high-rent locations. Yet exactly the opposite was seen to be occurring in Graves (1983), where migrants were moving *toward* high-rent locations, holding income constant. These two results strongly support a model in which the rising national income of earlier years led to greater demands for desirable locations, driving down wages and driving up rents, and in the 1960s, with income continuing to rise, migration continued to the desirable locations, despite the lower wages and higher rents observed in such locations.

Finally, note too that, as with the case of the “fast lane” on the freeway, it does not necessarily take many movers to yield an equilibrium in which real utility is approximately equal across space. This is not to argue that there are no longer *any* variations in utility over space, but rather that the dominant observed pattern is one of equilibrium (as discussed in Mueser and Graves 1995, shocks to employment continue to occur, but they are more intertemporally and spatially random than are the systematic amenity influences). To a large extent, the importance of demand-side influences relative to supply-side influences depends on the time perspective of interest: for near-term interests, the demand-side approach becomes more relevant (see Greenwood and Hunt 1989), while for longer-term interests, the supply-side approach becomes more relevant (as with the early supply-side approach of Borts and Stein, though that was driven by excess labor in agriculture rather than amenities).

2.4 Spatial Labor Market Equilibrium in the Urban/Regional View with Suggestions for Future Research

Taking the equilibrium view, the central observation to make is that there is no compelling reason to expect wage convergence. Indeed, since wage differentials are compensatory for amenity differentials, one might expect wage *divergence* over time, if desirable locations are also normal or superior, as might be expected. That is, as first discussed in Graves 1979, rising income nationwide will increase the demand for many things such as restaurant meals, clothing, automobiles, and the like – but while those goods can be incremented *in situ*, increased demand for lower humidity or more sunshine requires migration toward areas offering these non-tradable goods. Thus, one might expect that the ongoing migration from

undesirable (high-wage) locations to desirable (low-wage) locations would lead to a growing wage gap over time, not a narrowing gap as would be expected in the disequilibrium view.

On the other hand, rents might capture increasing percentages of the value of a location's amenities over time, as nice places become larger. This is particularly likely if, as will occur with an aging population, an increasing percentage of households have no members in the labor force – such households would be expected to move to locations where amenities are capitalized largely in wages, since they do not have to pay that compensation (see Graves and Waldman 1991). In the process of moving, these households will increase the share of amenity compensation occurring in land markets, at least to some extent. Moreover, if there are endogenous disamenities that are functions of city size (e.g., pollution or congestion), nice places might become less nice – the compensation for say good climate may go not entirely into wages and rents but also into offsetting endogenous “bads.” Hence, at the level of theory, it is unclear whether wage differentials over space would be expected to converge or diverge over time, when exogenous variables (e.g., national income growth and increasing average age) affecting the demand for locations change in magnitude.

However, dropping the assumptions made earlier that households have identical preferences and firms have identical production functions allows a fairly wide range of predictions regarding the spatial labor market equilibrium.

First, as already mentioned, an aging population with fewer labor force members will result in a higher percentage of households not in the labor force. The movement of these households to desirable locations with their amenity values largely capitalized into labor markets will drive up rents, reducing the percentage of the amenity value capitalized into lower wages. Similarly, such households leaving the undesirable locations will lower rents, decreasing the wage compensation necessary to equilibrate utility over space; hence, the mere fact that a population is aging results in wage convergence, a prediction that, as far as I know, has not been made before despite it being a fairly clear implication of the model.

More generally, the rich (high skilled and well educated) will outbid the poor (low skilled and poorly educated) for the desirable areas, much as they outbid them for BMWs, while the poor will outbid the rich for the undesirable areas, much as they outbid them for a 1980s K-cars. However, the rich in the desirable areas will demand the services of the poor. Since the rents will be determined by the rich buyers, the poor will be unwilling to work in the nice locations unless they receive wage compensation. The nature of the compensation will depend on how close desirable and undesirable areas are to one another. If they are quite close (e.g., as in some parts of Los Angeles), the necessary wage compensation will be the commuting cost of the poor. If there are no undesirable locations near to the desirable area, the necessary wage compensation will be the difference between value of the amenity to the rich and to the poor, adjusted for differential lot sizes for the two groups (a “stand-alone” topic we shall return to). In terms of the spatial labor market equilibrium, the wages of the poor will be *higher* in the desirable location – to those not carefully considering the situation, it might be inferred that the

desirable location to the rich is actually *undesirable* to the poor. This is another example of the tricky interaction between wage and rent compensation in the equilibrium urban/regional approach when heterogeneity of preferences is allowed.

In addition to aging and income, another exogenous variable with potentially important – yet unexplored – implications for the wage-rent hedonic analyses is the presence and number of children in a household. The effects of children are clear within an urban area – young married couples tend to centrally locate (to minimize average commute times and take advantage of central amenities such as restaurant and cultural diversity) until their children reach school age, at which point most move to the suburbs to obtain larger lot sizes, better education, and lower crime rates.

These intra-metro effects are likely to exist over a broader array of spatial locations, with larger families moving to metropolitan areas with lower housing prices vis-à-vis childless or small families. Those movements will have had impacts on the equilibrium wage compensation in the USA since the trend in average family size has been markedly downward from the 1950s, with 3.37 persons per household, to the present 2.6 persons per household (<http://www.infoplease.com/ipa/A0884238.html>). If the growing numbers of childless and small family households prefer high-amenity and more central locations, such locations will become more costly in land markets. Whether this leads to a lower or higher amount of amenity capitalization in labor markets depends on whether such families have a higher or lower number of labor force participants and housing density. If, as was the historical case, bigger families are likely to have fewer labor force participants as one spouse stays home to take care of the children, an influx of childless and small families might lower wage rates, leading to increasing divergence in wages over space.

The preceding examples of individual traits that vary among households (income, age, and number of children) are traits that are widespread in the population. This would lead to the expectation that, in equilibrium, utility will be equilibrated over space. That is, there will be no “spatial consumer surplus.” Essentially, at a full hedonic equilibrium, households could flip a coin to decide where to live, because compensation for variation in amenities would result in equal utility in all locations.

For some traits, however, it may well be that the number of households possessing a strong demand for a particular amenity is smaller than the number of locations offering that amenity. In this case, individual households can obtain spatial consumer surplus, being better off – possibly much better off – in some locations than in others. A disabled person, for example, might get far greater than average benefit from access to public transit, but the number of disabled individuals might be quite small relative to the number of locations offering that access. A passionate mountain climber might well obtain greater than normal satisfaction from occupying a town near climbing opportunities than do other occupants of that town. To the extent that unusual preferences relative to the opportunities available are important, amenities will be undervalued in land and labor markets by the hedonic method. It is quite likely that those in the “upper tails” of the demands for a wide variety of amenities might be paying less than their willingness to pay for the

levels consumed. The true value of the amenity, for example, public transit, is the sum of the observed willingness to pay plus any *unobserved* consumer surplus. Similarly, those in the “lower tails” of demands may receive more compensation than necessary for goods, also resulting in spatial surplus – if, for example, there is a high probability of death in a high-risk location, rents will be very low; one who does not care much about such risks achieves spatial consumer surplus by locating in such areas.

A long-standing interest in labor economics is the return to education. At first blush, it would seem that amenity compensation in labor markets would result in an *understatement* of the returns to education. Since the more highly educated would have higher lifetime expected incomes, regardless of location, one would surmise that they would want to locate in the more desirable locations, but since the desirable locations offer lower wages, *ceteris paribus*, the highly educated would appear to get a lower financial return from their education, since they would be taking part of that return in the form of amenity consumption. It turns out, surprisingly, that this is not the case (see Graves et al. (1999a) for more detailed discussion and a graphical treatment). The reason is that, in the actual array of locations in the USA, desirable locations for households are even *more desirable* to firms. Consider the upper left cell in Fig. 2.1, where it was shown that locations desirable to both firms and households will be large with high rents, while the impact on wages is ambiguous, depending on relative desirability which is an empirical issue. In the data of Blomquist, Berger, and Hoehn (and very likely most other more recent empirical studies), the locations that were desirable to households were even more desirable to firms; hence, while the greater supply of labor leads to wage reduction, the demand for labor is greater yet, leading on net to higher wages in such areas. Hence, earnings functions that aim to estimate the return to education *overstate* the benefits of education in analyses ignoring amenities (for a somewhat different approach yielding the same conclusion, see Dahl 2002).

Another long-standing issue in the labor economics literature is the magnitude of the return to unionization. Unions have, as a historical fact, been concentrated in the Northeast and upper Midwest. As noted in Graves, Arthur, and Sexton (1999a), all studies of unionization fail to control for amenities. In one of their specifications, fully one-half of the presumed benefits of unionization were seen to be related to the fact that unions were stronger in areas of less desirable amenities, particularly climate – unions are getting “credit” for what is really compensation for a bad weather. A detailed analysis at a more disaggregated level would be better able to separate the relative importance of amenities and unionization.

Interestingly, amenities are actually substantially more important than they appear in existing empirical studies, because these studies ignore fringe benefits due to limited data. It turns out that fringe benefits are spatially varying in ways that reinforce observed wage compensation for amenities (see Graves et al. 1999b). Fringe benefits are substantially higher in the Midwest and Northeast than they are in the South and West, perhaps in part because of structural differences in the nature of occupations among the regions. Hence, the higher wages that are paid in the

former regions to compensate for undesirable climates would be higher yet, were full compensation employed rather than just wage compensation. Similarly, the desirable South and West regions have both lower wages and lower levels of fringe benefits. If foreign and other competition is causing the fringe benefits to decline, as appears to be the case in the Northeast and Midwest, this would lead to wage *divergence* as the necessary compensation would cause wages to rise as fringe benefits fall in equilibrium.

As mentioned earlier, some authors have regarded the ratio of net migration to the gross flows as a measure of “migration efficiency.” In the context of the labor market disequilibrium approach, this notion makes a fair amount of sense – it would seem inefficient to have large numbers of people moving both in and out, when net in-migration is occurring. If people are moving in because wages are higher in a location, it would seem odd (“inefficient”) that many people are also moving out. Yet, an empirical regularity is that when net in-migration is large to a location, so are the flows of out-migration. In the urban/regional equilibrium view, this empirical regularity is actually to be expected. As individuals move in to, for example, desirable locations, they drive up rents and lower wages (and also increase endogenous levels of disamenities), which in turn results in others leaving, as an optimal reaction to these changes, not as a matter of “inefficiency.” Some will cash out of their houses as their property values increase resulting in a nonoptimally large share of wealth in housing. Others will leave as the property comes to be worth less to them than to the newcomers. And still others will leave because congestion and air pollution are of particular importance to them. Finally, some will leave because their wages are lower in ways that the amenity level no longer compensates for.

Another issue in the urban/regional approach, which has implications for spatial equilibrium in the labor market, is the appropriate capitalization rate to use when converting rents into property values or vice versa. Linneman (1980) and Linneman and Voith (1991) argue that to consider either rents or property values separately in a hedonic valuation function results in selectivity bias; hence, they should be considered together. However, doing so raises the question of how to merge rent data with property value data. In the earlier paper, Linneman found that a 3 % capitalization rate was appropriate to convert property values into rental flow equivalents for 1973 Chicago data. In the later study, a capitalization rate (varying with traits of the household head) was argued to be 10 % for 1982 data from Philadelphia.

For analyses within any particular housing market, it seems important to correctly merge the data to avoid selectivity bias present in using either property values or rents separately. If, however, a study is being conducted using data at a national or large regional level (to, e.g., estimate the value of a greater variety of amenity bundles), there are additional concerns. In areas expected to grow (in either size or value due to growing demand for the amenities offered), property values will be high relative to *current* rents, because those rents are expected to be increasing – there is the expectation of two forms of return to housing in growing areas, rents collected currently and growth in property value over time as the rental stream gets larger. Conversely, in areas expected to lose population, rents will be expected to

fall in the future, so current property values will be low relative to current rents, since a fall in values is expected. These results are required to have housing investment profitability be the same in both growing and declining markets.

To get a sense of the disparity in rent/value ratios, using 2009 census data, the entire state of Colorado had a median housing value of \$234,100 and a median monthly rental housing cost of \$835, for a rent/value ratio of .00356, or .04272, multiplying by 12 to annualize this ratio, for easier intuition and to compare returns to other assets. The state of Michigan had median housing value of \$147,500 and a median monthly rental housing cost of \$709, an annualized rent/value ratio of .05772. There is, perhaps not surprisingly, great variation of these numbers within states, and that variation is consistent with the arguments made here. For example, Aspen City (\$860,000, \$1,319,.01836) and Boulder City (\$464,200, \$998,.0258) have very low annualized rent/value ratios relative to Colorado as a whole, while Birmingham City (\$388,800, \$1,145,.03528) and Ann Arbor (\$244,300, \$950,.04668) also have lower numbers than averages for Michigan.

What are the implications of the preceding for the spatial labor market equilibrium? For locations that are expected to either grow in size or that possess superior amenities that are expected to be valued more in the future, using a single capitalization rate results in hedonic analyses that are biased. The US average annualized rent/value ratio is .05292 (\$185,400, \$817), while Hawaii's rent/value ratio is .02808 (\$521,500, \$1,221), and Oklahoma's rent/value ratio is .07452 (\$98,800, \$614) – these are the current annual returns (2.8 % and 7.5 %) necessary to have housing investments be equally profitable in both locations, allowing for expectations of growth and decline in rental returns, respectively. The percentage owner occupied in Hawaii is 58.1 % compared to Oklahoma's 67.9 % and a national average of 66.9 %. If the national capitalization rate were applied to Hawaii, imputed rents would be \$2,300/month, when actual rents were only \$1,221. Averaging the numbers with a weighting of 58.1 % on the former would imply a weighted hedonic rent-equivalent value of \$1,848, rather than the actual rental rate of \$1,221; for Oklahoma, using the national capitalization rate would result in a weighted hedonic rent-equivalent value of \$493, far below the actual \$614 rents actually being paid. Hence, using a single capitalization rate in a national hedonic study will bias *upward* the rents estimated for nice locations and will bias *downward* the rents estimated in more undesirable locations – if, on the other hand, rents were capitalized up to property values with a constant national capitalization rate, property values would be biased *downward* in nice places and biased *upward* in less nice places.

Assuming that property values are converted to rents, and under strong homogeneity assumptions that rental housing and owner-occupied housing are equivalent, as are renters and owners, then labor would “look” from the hedonic housing models to require less compensation in nice places (since more of the niceness appears to be going into rents than is actually the case) and to require more compensation in the undesirable places because less disamenity appears to be capitalized into rents in those locations. In light of the difficulties raised here, along with the likelihood that rental housing and owner housing are different as

are renters and owners, it would seem that an argument could be made for conducting separate analyses for each group, resulting in different amenity values for each group. Obtaining the “true” amenity value of a location, then, might merely be a matter of weighting the values obtained in the separate analyses by the percentages of people in each group, which would vary by location.

Closely related to the preceding difficulty with hedonic models is the ubiquitous assumption in the theory of a constant lot size and a constant dollop of work effort (the 40-h week), each normalized to unity. This would not seem, at first thought, to be a great difficulty at the empirical level since the labor hedonic data could be restricted to full-time workers and the housing hedonic could include lot size as an explanatory variable. However, both the quantity supplied (e.g., perhaps fewer hours at lower wage rates in nice places) and the supply of labor (e.g., shifting if leisure is a complement or a substitute with amenities) are likely to vary in what are currently unknown ways with wage variation due to variation in amenities. Moreover, any particular wage level can occur with either high rents (if a location is high in either household amenities or firm amenities or both) or low rents (if a location is undesirable to either or both) – and one would generally expect that housing prices would not be independent of work effort, apart from simple Cobb-Douglas utility characterizations. If leisure is complementary with amenities, the supply of labor will be lower in nice places (wages higher) and higher in undesirable locations (wages lower). The assumption of a fixed amount of labor in all locations will then bias downward the apparent value of amenities under complementarity. Moreover, if desirable locations are also superior, the assumption of a constant amount of work effort over space will, then, result in a bias that will, over time, look like more convergence is going on than actually is.

In addition, how to handle lot size is complicated. Consider an amenity bundle, common in practice, which is comprised of amenities whose consumption is independent of lot size – for example, access to the central business district in the standard urban model or access to a wide variety of other amenities, such as nearness to an ocean or the breathing of air of various quality levels. In such situations, one would expect substitution of capital for land to occur (e.g., high-rise buildings as one approaches the CBD radially). How much is being paid for the amenity in this case depends critically on lot size – if one buys twice the average lot size, one is paying twice as much as others for the amenity. This implies that merely holding lot size constant in the rent hedonic is insufficient; to obtain marginal prices, an interaction term between lot size and the various amenities must be introduced.

If actual lot sizes get smaller in high-amenity locations, as would generally be expected, the assumption that lot sizes are constant leads to a bias that underestimates the amenity values. And, if smaller lot sizes, *ceteris paribus*, are less desirable than larger lot sizes, again as expected, the nice places are a little less nice for this reason; hence, wages in nice areas would be biased *upward* by the constant lot size assumption, while wages would be biased *downward* in the less desirable areas where lot sizes would be larger than average. Thus, there appears to be greater convergence in wages than there truly is, just because of the assumption of constant lot size.

The standard models also assume competitive land and labor markets. As but one important case where this assumption is not valid, consider the California Coastal Commission that regulates building construction in coastal California. Were it not for the stringent zoning of this commission, it is very likely that virtually the entire coastline of California would look like Collins Blvd in South Beach, Miami, with high-rises lining the ocean and extending inward. This might result in a much larger percentage of the US population living in California. The “value of ocean access” would be seen to be vastly higher in such a world, aggregating over the many consumers. This is not necessarily to argue that the zoning is inefficient as it is possible, though I suspect highly unlikely that nonuse values of all Americans might exceed the use values of the many millions of residents who would occupy those buildings. The scenic views from the Pacific Coast Highway certainly have value, to Californians and visitors alike, but those values are not being captured by property value studies, since the properties that “would” be there in a free-market setting are prohibited by the Coastal Commission. The large-lot zoning requirements effectively restrict ocean access to the very rich (e.g., as in Malibu) who are willing to pay a great deal for ocean access with the less rich who would like to acquire ocean access along with potentially much smaller lot sizes being effectively excluded by CCC zoning laws.

Taxation at the federal, state, and local levels also modifies the competitive outcome. Progressive federal taxes will tend to encourage movement toward locations in which amenities are capitalized into lower wages, since a lower percentage tax on true income can be obtained in such locations. To the extent that state income taxes paid are not closely related to services received, they too will distort the location decision, the empirical question being how are state income tax levels related to amenities available by state? Variation in property taxes similarly will distort location decisions, with interesting effects depending on the correlation between amenity levels and property tax rates. Even varying sales taxes over space will have impacts since local non-tradable goods will be less expensive in locations in which amenities are more capitalized into wages. All of such tax effects have been little studied as far as I know yet could yield important insights – for example, if rents are high for firm productivity reasons, a high property tax will be compensated for in equilibrium with higher wages, but if rents are high due to household amenities, the higher property taxes would just be another portion of the price paid for amenities in land markets, resulting in higher wages than would otherwise be the case.

The hedonic method implicitly assumes that all amenities associated with a location are accurately perceived by households and firms. This is not controversial for amenities whose benefits are sensed by our five senses (e.g., view premiums, the sound of the ocean, smells of various sorts, the feel of warmth on the skin, the diversity of tastes available in locations with many fine restaurants). However, there are amenities whose benefits are unlikely to be fully captured by the senses. Environmental improvements, for example, might be partially perceived by the senses, but complex health effects, the magnitude of which experts in the field argue about, are unlikely to be perceived accurately if at all. In such cases, the hedonic

method is very likely to undervalue the amenity, with property values too low and wage rates too high in the clean locations.

Some effects might even be quite misperceived. For example, acid-polluted lakes offer much greater water visibility than do non-acid-polluted lakes – cleaning up such lakes might lead property values to fall around them, if people think that being able to see deeper in the lake is an important trait. Nearness to hazardous waste dumps has a very large negative effect on property values (and perhaps wages, to the extent that wage variation occurs within labor markets, as seen in Blomquist, Berger, and Hoehn), even when knowledgeable experts assert that there can be no local effects associated with the dump. Individuals receive more radiation leaning against a granite wall in Grand Central Station than they would receive leaning against the outside wall of a nuclear reactor, yet thousands do the former every day that would be horrified to contemplate the latter.

What is one to make of these examples? In the case of amenity benefits that are not fully perceived, an argument could be made for adding benefits from health effects models (e.g., number of asthma attacks averted times the willingness to pay for an averted asthma attack, number of lives saved times \$7 million dollars, the current value of a statistical life saved being employed by the EPA) to those from hedonic models. However, this is likely to involve some double counting as households might infer that smelly air is unhealthy air. Also, if an individual “feels” damaged by a hazardous waste dump or a nuclear reactor, then is that not a *real* damage? If that individual gets an ulcer from worry, it is still an ulcer. One might argue that public authorities should attempt to educate households about the true risks of damage they face from various sources, since households are notoriously bad at assessing such risks. On the other hand, the dread associated with some risks (e.g., cancer, terrorist attack) may truly be greater than that associated with other risks with similar “outcomes” (e.g., dying in a car crash), and willingness to pay to avoid the first group of risks may be genuinely much greater than WTP for the latter risks.

It should be noted that the array of amenity levels among locations is not independent of either technology or public policy. The creation of the interstate highway system in the 1950s, the 1960s, and continuing into the present has hastened the decline of the Rustbelt and the expansion of the Sunbelt. However, the latter expansion would have been much smaller were it not for the invention and widespread innovation of air conditioning in the South and Southwest. Uniform national environmental standards (e.g., the requirement that all cars be equipped with catalytic convertors) have the practical effect of causing movement to the areas that most benefit from such policies – Los Angeles, with frequent stagnant air conditions, benefits more from such policies than does Chicago. What these examples imply is that one cannot run a hedonic equation at one point in time and apply the results to time periods far before or far after that study.

In certain relatively rare cases, the nature of the underlying preferences for an amenity matters greatly to its valuation. Normally, economists do not care at all “why” households desire the goods they desire, it not mattering whether one person wants a refrigerator to keep beer cold, while another wants a refrigerator for fresh

produce or ice cubes. In either case, the estimation of the price, cross-price, and income elasticities of interest to the economists is unaffected. Even in situations in which economists think about the underlying motives (as with the medium of exchange, asset, and precautionary motives for holding money), the estimations and conclusions are unaffected by those thoughts. For environmental goods, however, the nature of the preferences matters in a way not widely known, as suggested by the California Coastal Commission discussion above.

Environmental economists typically talk about (i) use values, (ii) option to use values, (iii) bequest motives, and (iv) preservation/existence values. Unlike the case of the refrigerator, these values frequently “clash,” in the sense that some households want to use an amenity directly, while others would like to preserve the amenity in its unused state. Are the demands for nonuse of the California coastline as large as or larger than the use values? Is Central Park in New York City more valuable as a park than the billions, perhaps trillions, of dollars it would be worth if developed? Is it better or worse to allow snowmobiles in Yellowstone Park in the winter when their noise and pollution disturbs the wildlife at a time when other stresses on the animals are at their annual peak? These are difficult questions, yet decisions have to be made; the decision to do nothing is itself a decision with costs and benefits. The decisions in these clashing cases are difficult largely because there is great controversy about the methods of ascertaining nonuse value vis-à-vis the methods employed – one of which is the hedonic method discussed here – to get estimates of use values. The takeaway message, though, is that the nonuse value of the amenity, from society’s perspective, *might* be larger than the benefits of using the amenity, the latter being measured by the higher property values and lower wage rates associated with using the amenity.

The discussion of this section has involved many topics related to the spatial labor market equilibrium. Many of these topics are either not discussed at all in the existing literature or the discussions are, as here, unduly preliminary to obtain solid policy-relevant conclusions. It is to be hoped, however, that the research initiatives sketched in this section will lead to more substantive contributions in the years to come.

2.5 Conclusions

In this chapter, two quite contrasting views of the nature of spatial labor markets have been examined. The notion that variation in wage rates represents variation in utility levels was appealing when costs of movement were high and when information about the nature of alternative locations was low. Those assumptions are increasingly irrelevant to observed movement patterns, and the bulk of this chapter took a polar opposite approach – assuming that variations in wages (and rents) occur as compensation for variation in amenities over space. In this latter view, expectations about wage convergence become quite ambiguous, depending on a wide variety of factors, many touched on in the previous section. It is argued here that the urban/regional view of the spatial labor market equilibrium is of

growing relative importance in the understanding of labor markets in the United States, and this is likely to be the case in the rest of the world in the years to come. Future research efforts pursuing in more detail the somewhat speculative assertions made throughout this chapter are likely to have important payoffs in terms of advancing our knowledge of labor markets and how those interact with land markets and amenity variables.

References

- Bloomquist GC, Berger MC, Hoehn JP (1988) New estimates of the quality of life in urban areas. *Am Econ Rev* 78(1):89–107
- Borts G, Stein J (1964) Economic growth in a free market. Columbia University Press, New York
- Chiswick BR, Hatton TJ (2002) International migration and the integration of labor markets. Discussion Paper No. 559, IZA
- Dahl GB (2002) Mobility and the return to education: testing a Roy model with multiple markets. *Econometrica* 70(6):2367–2420
- Graves PE (1979) A life-cycle empirical analysis of migration and climate, by race. *J Urban Econ* 6(2):135–147
- Graves PE (1983) Migration with a composite amenity: the role of rents. *J Reg Sci* 23(4):541–546
- Graves PE, Linneman PD (1979) Household migration: theoretical and empirical results. *J Urban Econ* 6(3):383–404
- Graves PE, Waldman D (1991) Multimarket amenity compensation and the behavior of the elderly. *Am Econ Rev* 81(5):1374–1381
- Graves PE, Arthur M, Sexton RL (1999a) Amenities and the labor earnings function. *J Labor Res* 20(3):367–376
- Graves PE, Arthur M, Sexton RL (1999b) Amenities and fringe benefits: omitted variable bias. *Am J Econ Sociol* 58(3):399–404
- Greenwood MJ, Hunt GL (1989) Jobs versus amenities in the analysis of metropolitan migration. *J Urban Econ* 25(1):1–16
- <http://factfinder2.census.gov/faces/nav/jsf/pages/index.xhtml> (source for rent/value data).
- Linneman PD (1980) Some empirical results on the nature of the price function for the urban housing market. *J Urban Econ* 8(1):47–68
- Linneman PD, Voith R (1991) Housing price functions and ownership capitalization rates. *J Urban Econ* 30(1):100–111
- Mueser PR, Graves PE (1995) Examining the role of economic opportunity and amenities in explaining population redistribution. *J Urban Econ* 37(1):1–25
- Partridge M (2010) The dueling models: NEG vs amenity migration in explaining U.S. engines of growth. *Papers Reg Sci* 89(3):513–536
- Roback J (1982) Wages, rents, and the quality of life. *J Polit Econ* 90(6):1275–1278
- Rosenbloom JL (2002) Looking for work, searching for workers: American labor markets during industrialization. Cambridge University Press, Cambridge
- Sjaastad LA (1962) The costs and returns to human migration. *J Polit Econ* 70(5):80–93
- Taylor LO (2008) Theoretical foundations and empirical developments in hedonic modeling. In: Baranzini A, Ramirez J, Schaefer C, Thalmann P (eds) Hedonic methods in housing markets: pricing environmental amenities and segregation. Springer Science + Business Media LLC, New York

Labor Market Theory and Models

3

Stephan J. Goetz

Contents

3.1	Introduction	36
3.2	Labor Market Theory	37
3.2.1	Labor Supply	37
3.2.2	Labor Demand	43
3.2.3	Labor Market Equilibrium	45
3.3	Defining Labor Market Areas	47
3.3.1	Historical Efforts	48
3.3.2	Cluster-Based Analysis	49
3.4	Labor Market Area Analyses	52
3.4.1	Explaining Differences in Labor Earnings	52
3.4.2	Models of Spatial Adjustment: Booms and Busts	53
3.5	Conclusions	55
	References	56

Abstract

This chapter reviews labor supply, demand, and equilibrium topics with the goal of showing how they determine labor market area (LMA) outcomes across geographic space. Labor supply curves are based on utility-maximizing choices between working and leisure, subject to a budget constraint, while labor demand curves are derived from the firm's production function assuming profit-maximizing behavior. The challenges of defining and empirically delimiting LMAs are examined from historical perspectives and using statistical clustering analysis, with commuting data serving as a key tool. A key distinction is drawn between functional versus homogenous regionalization problems, and a number

S.J. Goetz

Northeast Regional Center for Rural Development and Department of Agricultural Economics,
Sociology and Education, Pennsylvania State University, University Park, PA, USA
e-mail: sgoetz@psu.edu

of suitable statistical approaches are reviewed. Current models used to study differences in earnings across labor markets as well as the effects of boom and bust cycles are also discussed. An empirical technique is presented for decomposing employment change within a community into four key labor market concepts: commuting, unemployment, labor force participation, and migration.

3.1 Introduction

The highest average wages earned in any US county in 2009 (\$90,500 in New York, NY) were nearly nine times higher than the lowest wages (\$11,400, in Worth County, Missouri) (Bureau of Economic Analysis, Regional Economic Information System). Across the NUTS 1 (Nomenclature of Territorial Units for Statistics) regions of the EU, average gross annual earnings in industry and services in 2006 ranged from EUR 72,038 in UKI London, compared with only EUR 2,397 in BGS Severna/Iztochna Bulgaria (Eurostat, [earns_ser06_26](#)). Understanding the reasons for these vast discrepancies across regional labor markets is a central objective of this chapter. More specifically, this chapter presents theory and models used by economists and regional scientists to analyze and understand spatially varying labor market variables including labor supply and demand, wages and productivity, and employment or unemployment, along with changes in these variables over time. The discrepancy in wages across US counties and EU regions provides a first important indication that labor markets do not simply equilibrate wages over space as might be expected, for example, in the case of the price of apples net of transport costs. With perfect knowledge and foresight and *all else equal* including the distribution of worker skills, labor (or firms) would migrate in response to wage differentials until the price of labor was the same in different locations. Instead, persistent wage differentials and variation in unemployment rates over space suggest that rigidities and other factors play important roles in the labor market that are worth studying. Of course, the average quality of labor as measured by skills also varies across labor markets, but this raises the question of why some markets attract higher shares of skilled labor than others.

This chapter is organized as follows. After reviewing basic microeconomic labor theory including labor force participation and discussing how labor market areas have been defined in the literature, applied models that have been used to study wage differences across labor market areas are examined. Chief among these are models of the returns to education and recent studies that examine differences in labor productivity due to spatial agglomeration and clustering. The economics of agglomeration has become a well-established area within regional science and labor economics, but it is receiving renewed attention with growing concerns about wage inequality (e.g., Goetz et al. 2011) and, as the world becomes more urbanized, with some cities attracting more economic growth than others (e.g., Glaeser 2008, 2011).

Although their role in determining local labor market outcomes is not explicitly discussed here, it is important to note that institutions such as labor unions

also matter. To the extent that unionized workers have higher earnings and more generous benefits, labor bargaining associations can produce different outcomes in local labor markets for equivalent labor efforts. When the Boeing aircraft company announced in 2011 that it would manufacture some of its new 787 Dreamliners in South Carolina rather than the State of Washington, it was accused of union “busting” and retaliation against unionized workers at the Seattle plant by paying lower wages to nonunionized laborers in the South.

Important labor topics including unemployment, migration, commuting, spatial mismatch, and job search are discussed in other chapters within this volume, as is spatial equilibrium in labor and housing markets. Nevertheless, this chapter briefly ties these topics together empirically in Sect. 3.4 on labor market models, where localized economic booms and busts are also discussed. The next section starts by outlining the basic neoclassical model of a labor market.

3.2 Labor Market Theory

To understand regional or spatial variations in basic employment measures, we start with the microeconomic determinants of labor supply and demand in an *aspatial* context. Using this framework, the rich and interesting causes of different labor market outcomes can be studied along with potential implications for policy and further research. This section builds on Cahuc and Zylberberg (2004), who do not discuss local, spatial, or regional labor markets or list these topics in their index. In fact, standard micro- and macroeconomics textbooks ignore or abstract from the effect of space altogether, and this is also true in the area of labor economics.

3.2.1 Labor Supply

A worker’s decision to supply labor to the job market is the result of an optimal choice between earning income (y), which enables consumption (c), and leisure (l) time. In the basic model, the individual worker faces a time constraint of 24 h a day, exogenous income ($y_g \geq 0$), which could be an inheritance or a spouse’s earnings, and wage rate w . The only decision to be made is how much leisure time to take given the wage rate, and this is given by the tangency between the worker’s indifference curve (u) and budget constraint ($c - wh - y_g \leq 0$) that is determined by income. After subtracting hours of leisure time from the 24 total hours available, we are left with the number of hours worked (h) at the given wage rate (w), which provides earned income.

This decision problem is shown graphically in Fig. 3.1 where total income is measured in the upper panel along the vertical axis and leisure time is recorded along the horizontal axis. Also shown are the utility curves ($u_1 > u_0$) that trace out the worker’s points of indifference between leisure (or working) and income. These curves have the usual properties of being convex to the origin, to reflect a diminishing marginal rate of substitution between leisure and income

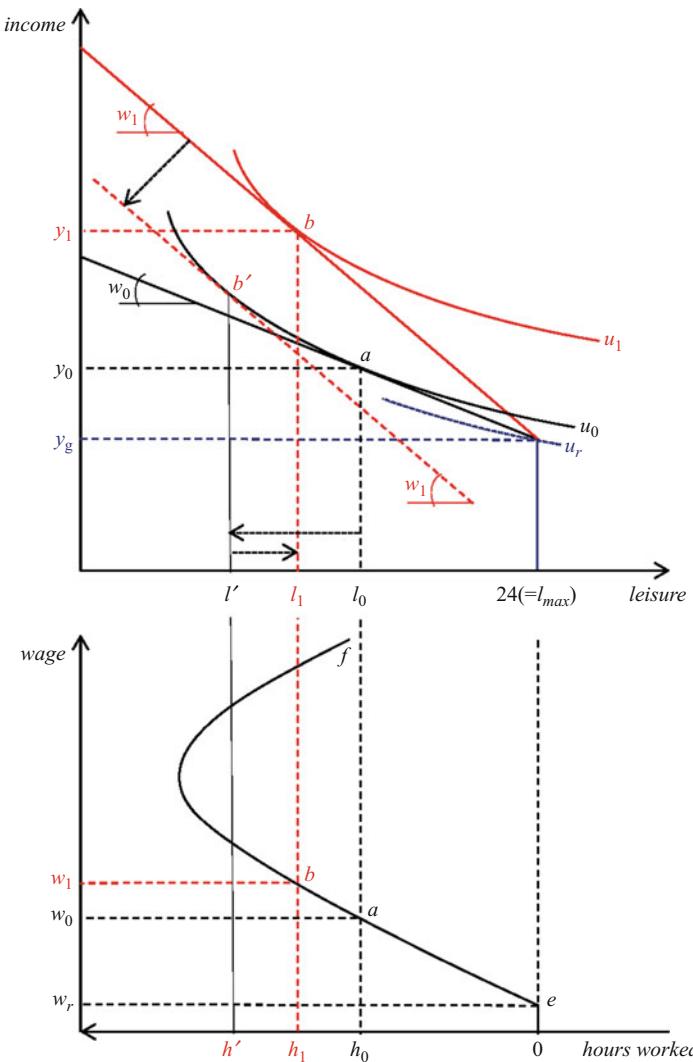


Fig. 3.1 Derivation of the labor supply curve for an individual worker

(or consumption), and they are continuous and twice differentiable. The budget constraint is shown as the unlabelled straight line, which has been shifted vertically by the amount of exogenous income (y_g) and has a slope given by the wage rate:

$$y = y_g + w_0 h (= c) \quad (3.1)$$

where $h = (24 - l)$ so that $dh/dl = -1$. For each additional hour of work or leisure foregone, income rises by w_0 (or w_1). Hence, the opportunity cost of an hour of leisure is w_0 .

An initial equilibrium (optimum) may be given at point a which provides the highest possible utility given wages of w_0 ; here, the worker consumes l_0 units of leisure and y_0 worth of goods while supplying h_0 h of labor. As the wage rate rises from w_0 to w_1 , the worker reduces the amount of leisure time taken and works more hours. This new optimal point is b , and it is determined by the tangency between the higher indifference curve u_1 and the new budget line reflecting the higher wage. The clockwise rotation in the budget constraint brought about by the higher wage relaxes the income constraint and allows the worker to reach a higher level of utility.

The lower panel of Fig. 3.1 shows the number of hours worked versus wages earned (with the x -axis inverted, i.e., increasing hours worked are measured from right to left). The individual worker's labor supply curve is given by the line $0ef$. The graph shows an area in which the supply curve is backward bending: Wages can eventually rise so high as to lead to a reduction in hours worked. In other words, when the marginal utility of income falls below the marginal utility of leisure, then leisure is chosen over work. Of course, this assumes that workers can choose the actual number of hours worked each day, which in many cases is unrealistic. The lower panel in Fig. 3.1 shows another important concept, that of the reservation wage $w_r \geq 0$, which is the wage level below which a worker chooses not to supply his or her labor, resulting in nonparticipation. As we will see later, this wage has important implications for who benefits from different kinds of local development projects, among other outcomes. In fact, the worker's problem consists of two different parts: first, whether or not to work and, second, how much labor to supply, conditional on having decided to work. If $w < w_r$, then the worker will not supply any labor, preferring instead to consume at a corner solution on the indifference curve labeled u_r in the upper panel of Fig. 3.1. An increase in wages in a local labor market may therefore induce workers not only to supply more hours but also to cause more people to work, that is, to enter the workforce.

We can state the worker's problem formally as that of maximizing the utility of consumption (or income) and leisure subject to the budget constraint:

$$\max u(c, l)$$

$$\text{subject to } c \leq wh + y_g$$

Note that there is another implicit constraint involved, in that $h + l \leq 24$ h. Using a Lagrangian multiplier and the shadow price of leisure in the complementary slackness condition, this optimization problem is solved for an interior solution (i.e., $h > 0$) by differentiating with respect to h and equating the result with zero. This yields

$$\partial u / \partial l = w(\partial u / \partial c)$$

The individual reaches an optimum when the contribution of the last unit of leisure to utility equals the value of the last unit of consumption, multiplied by the wage rate. In other words, the wage rate at the optimum equals the ratio of the marginal utility of leisure to the marginal utility of consumption, which can be simplified to $w = (\partial c / \partial l)$. In Fig. 3.1, this means that we have reached a point of tangency between the budget constraint and the indifference curve; here, the marginal rate of substitution of leisure for consumption is equal to the wage rate. For example, this is illustrated by point a along indifference curve u_0 .

We can also show the decomposition of a wage change into income and substitution effects familiar from consumer theory. To do this, assume again that the worker has exogenous income in addition to wage earnings, so that as before, $y = wh + y_g$, and $h = h(w, y_g)$. Using the fact that $l = 24 - h$, we in this case can write that

$$\partial l / \partial w = (\partial l / \partial w)|_U + l(\partial l / \partial y) \quad (3.2)$$

Here, the effect of wages on leisure (or its inverse, hours worked) is decomposed into a substitution effect in which utility is held constant (i.e., we are moving along the indifference curve, u_1 , which gives the compensated, Hicksian labor supply) and a pure income effect (i.e., the budget constraint is shifted to the left by Δw , yielding the uncompensated, Marshallian labor supply which accounts for the income changes associated with the wage increase). In Fig. 3.1, the substitution effect occurs as a movement along u_0 from point a to b' . To get this point, we basically take away the income gained during the wage increase so as to keep the worker on the same indifference curve. This provides the (labor for leisure) substitution effect, which is always negative because at the higher wage the worker finds it worthwhile to work *more*. Subsequently, the income is restored which allows the consumer to reach the higher indifference curve (u_1), and this is the income effect. Now that the worker is better off, he or she can afford to work fewer hours, enjoying more leisure.

So long as leisure increases with income, it is a normal good. When the income effect exceeds the substitution effect, however, the labor supply curve can bend backward; in Fig. 3.1, point b could be located to the right of point a . Indeed, since at least the middle of the last century, men have been working fewer hours and participated less in labor markets, while women have entered the workforce in increasing numbers and they are also working more hours. To the extent that workers in different regions of a nation, that is, in different labor markets, make these trade-offs in different ways, it is already clear that labor supply varies over space and so will labor market outcomes.

As one interesting implication, Cahuc and Zylberberg (2004: 13) show that the reservation wage (w_r) satisfies the following equality:

$$u[y_g + w_r(24 - l_c), l_c] = u(y_g, 24) \quad (3.3)$$

Here, l_c is the constrained number of hours that the individual must work, in a take it or leave it full-time job situation. For example, this may be a mandatory 8-h workday, when in fact an individual may prefer to work only 6 h, which would

Table 3.1 Labor force participation rate (φ), November 2011, selected states, seasonally adjusted

Lowest-ranked states	Rate	Highest-ranked states	Rate
West Virginia	53.0	Vermont	70.8
Alabama	58.4	South Dakota	70.8
Louisiana	59.1	Minnesota	71.6
Michigan	59.8	Nebraska	71.8
New Mexico	60.0	North Dakota	73.2

Source: US Department of Labor, Bureau of Labor Statistics, Local Area Unemployment Statistics, December 20, 2011; estimates are based on population aged 16 years and older

allow him or her to reach a higher indifference curve. The worker is indifferent – or at the tipping point – between not working (living off exogenous income and consuming the maximum amount of leisure) and working the required amount of time. This is an important point, because a slight change in the wage will cause the individual to drop out of the workforce, causing nonparticipation.

Next, by summing the individual labor supplies (number of hours worked by each individual i), we can calculate the aggregate labor supply, $N_F = \sum h_i$, for a nation as well as for individual regions or labor market areas, as defined below. The labor force participation rate out of the total population (POP) is defined as $\varphi = LF/POP$, where the labor force $LF = N_F + N_U$ is comprised of the number of employed (N_F) and unemployed (N_U) workers. This also means that $v = N_U/LF$ is the unemployment rate, to be determined below. Note that the labor force participation rate is the number of individuals who are either working or unemployed and actively looking for work, as a share of working age individuals (usually 16–65 years of age or simply 16 and over) in the population. [Table 3.1](#) shows the five states with the highest and lowest levels of φ . The high rate of labor force participation in North Dakota, which is going through an economic boom related to oil exploration in this period, is noteworthy, as is the very low rate in West Virginia.

There also is a relatively systematic relationship between the labor force participation and unemployment rates. [Figure 3.2](#) shows a pronounced negative association between φ and v : About one-half of the variation in the participation rate across state labor market areas is associated with variation in unemployment. The labor market of North Dakota has both the lowest rates for v and the highest rate for φ , suggesting that the high local earnings associated with the oil boom and the resulting low local unemployment are drawing large shares of workers into the labor force.

The size of the labor force at any moment in time and in any given labor market area depends on the wage rate and, more specifically, the reservation wage w_r . To see this, consider a cumulative distribution function, $cdf(\cdot)$, also shown in [Fig. 3.3](#), which represents the distribution of reservation wages within the population; for example, this may be a standard normal distribution. Then, it is true that $cdf(w)$ measures the share of working age population for whom $w_r < w$, or the participation rate, and $POPcdf(w) = LF$. Furthermore, note that $\partial LF/\partial w > 0$ because the cdf is an increasing function. These relationships are graphed in [Fig. 3.3](#).

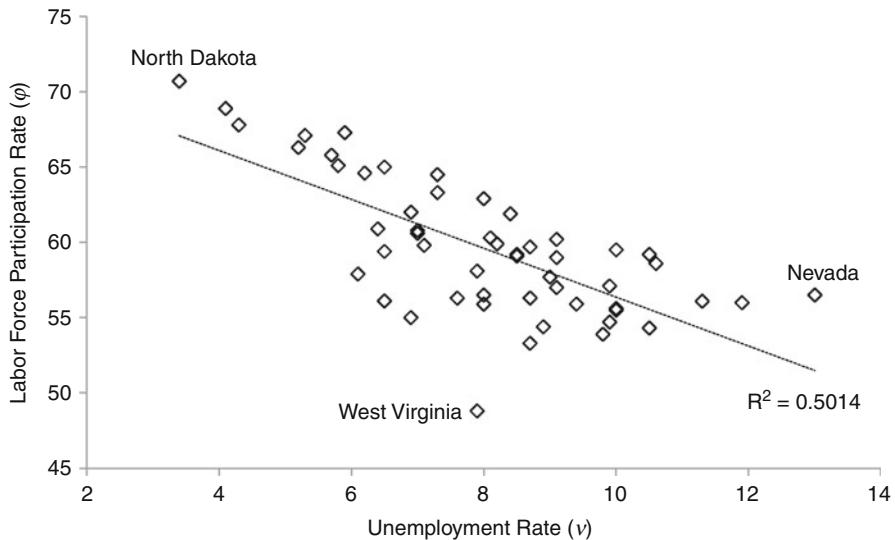


Fig. 3.2 Relationship between φ and v in state labor market areas, November 2011 (Data Source: Local Area Unemployment Statistics, BLS, January 2012)

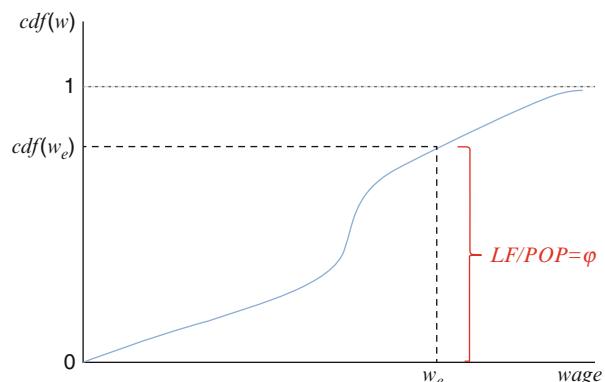


Fig. 3.3 Cumulative distribution function for wages

There is one other important group of workers relevant especially in regional and national labor market analyses, and that is the number of discouraged workers – those who have given up the search for work. These workers' reservation wages would be covered in the local market, but their employment prospects are so weak that they have given up the search for work and are no longer counted as being part of the *LF* (see also ► Chap. 4, “Job Search Theory” in this volume). It is important to consider these individuals in local analyses because they may jump back into the labor market if there are new jobs. In turn, this may lead to a short-run increase in unemployment rates, if a (new) factory starts to (re)hire workers locally. The aggregate number of discouraged workers in the USA was estimated to be

around one million in January 2012 (Bureau of Labor Statistics), and it can therefore not be ignored by policymakers. The share of discouraged workers in the labor force also varies over space, that is, by labor market area.

So far, we cannot say anything about unemployment because that depends also on labor demand. We return to this important concept briefly after discussing labor demand; a more thorough treatment can be found in ► Chap. 7, “[Regional Employment and Unemployment](#).” Here, the subject of job search is important as it relates to nonparticipation as well as the phenomenon of discouraged workers.

In the conventional (aspatial) textbook treatment of labor supply, other topics that are covered here include labor supply elasticities (with those of men being higher than those of women), human capital and the role of education in determining worker outcomes, and the returns to education. A number of models have been developed to study these topics, and empirical research has been carried out on varying returns to education across labor markets. We have also glossed over other important rigidities in the labor market, such as the transaction costs involved in finding work as well as working, including the need for a wardrobe and transportation, and in the decision to work in more than one job (i.e., multiple jobholding).

3.2.2 Labor Demand

Like other inputs, labor (n) is a factor of production for the firm and as such is subject to a derived demand. The production function $q = q(n, k)$, where k is a fixed capital input, is assumed to be strictly increasing and concave, so that the marginal product increases ($q' > 0$) at a diminishing rate ($q'' < 0$) over the relevant range. Assuming that labor, at wage w , and capital, at rental cost r , are the only inputs, the firm maximizes the following profit function with respect to the variable input, n :

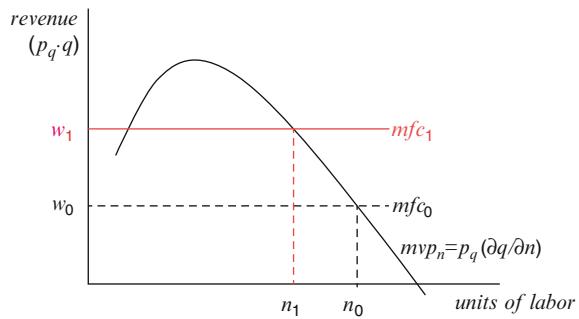
$$\max.\pi = pq(n) - wn - rk \quad (3.4)$$

Since $dk = 0$ in the short run, this yields $\partial\pi/\partial n = 0 = p\partial q/\partial n - w$ or $mpp_n = w/p$.

The firm’s demand for labor depends on a number of factors, including the marginal productivity of inputs (i.e., the curvature of the production function), and on relative prices of factors as well as output, which in turn depend on the firm’s market power. Graphically, the demand for labor is that segment of the marginal physical product (mpp) curve that lies in stage II of the static neoclassical production function, which is essentially a cubic function with an inflection point and a turning point (i.e., $q = a + bn + cn^2 + dn^3$, where $a \geq 0$, $c > 0$, and $d < 0$). Stage II is the area between the inflection point, or maximum mpp , and the point of maximum production, where $mpp = 0$.

[Figure 3.4](#) depicts the relationship between the firm’s revenue, or its output (q) scaled by the price of the output (p_q), and the amount of labor (n) used in production. The relevant decision range for the firm is the area between maximum

Fig. 3.4 Derivation of the firm's demand for labor



mpp (or mvp) and $mvp = 0$. As the cost of labor, or marginal factor cost $mfc = w$, rises and falls, the firm determines the amount of labor hired by the intersection point between the curves, that is, $mvp = w$, thus tracing out the conditional labor demand schedule, which is unambiguously downward sloping (because $\partial p_q q / \partial n < 0$). Hence, $n_1 < n_0$ since $w_1 > w_0$. If there is another factor of production, such as capital, then the degree to which labor demand is adjusted in response to wage changes depends also on the elasticity of substitution between the two factors.

One other variable important for determining labor demand is the output price, p_q . This price may vary across labor markets, for the same good, and different firms may also have varying degrees of market power locally, allowing them to set prices. To examine this effect, we start with the inverse demand function facing the firm, $p = p(q)$ with elasticity $\xi = qp'(q)/p(q) < 0$, which we also assume for the sake of simplicity to be constant.

The constant ξ has a few convenient properties, including that $\xi = 0$ under perfect competition, that is, when the firm takes market prices as given. When this condition no longer holds, that is, $\xi < 0$, the firm has some power to set prices by changing output levels. The larger is $|\xi|$, the greater is the firm's market power. Of course, ξ also depends on the actions of other firms, but if a particular chain has driven competitors out of local markets, that, too, would increase its market power in those markets. Finally, using the notation from Fig. 3.4 and the profit-maximizing equilibrium condition that $mfc = mvp$, we can show that (see also Cahuc and Zylberberg 2004: 175)

$$mpp_n = \rho(w/p_q) \quad (3.5)$$

where $\rho \equiv (1+\xi)^{-1}$ is a markup factor determined by the firm's market power. The firm's profits are maximized when mpp_n equals the markup factor ρ times the real wage paid by the firm. In the absence of market power, the elasticity is zero and the markup $\rho = 1$. A similar result is obtained by examining the cost of production, where it can be shown that at the profit maximum the marginal cost of production $c'(n)$ is multiplied by the markup factor: $p_q = \xi c'(n)$. And, under perfect competition ($\xi = 1$), the good is priced at precisely the marginal cost.

3.2.3 Labor Market Equilibrium

At last we turn to combining labor supply and labor demand factors to arrive at equilibrium in the labor market. Usually the topics of aggregate labor supply and demand in a nation are covered within macroeconomics and involve other variables such as aggregate production and demand and prices and inflation as well as fiscal and monetary policy instruments for stimulating the economy. As already noted, the subject of regional or local labor markets is usually not covered in economics textbooks, including Cahuc and Zylberberg (2004), although studies do exist of labor markets in different countries, including those making up the European Union.

Given a downward-sloping labor demand curve and an upward-sloping labor supply curve in wage-labor space, equilibrium wages and employment levels are determined at the point of intersection. This is shown as wages w_e and aggregate employment N_F in Fig. 3.5. In practice, there is always at least some unemployment in the economy, which in the graphic is shown as the (vertical) distance between the labor supply curve and the vertical line denoting the labor force (LF), which represents maximum feasible employment. Thus, at any given wage, we have that $N_U = LF - N_F$ or the number of unemployed people who would be willing to work if they could find a job. The supply curve asymptotically approaches the line LF showing that as wages rise, more and more workers see their reservation wage exceeded and wish to work, until the maximum is reached.

In this model, wages will rise, along with employment levels, in response to a demand shock. For example, a new firm may locate in a local labor market or the price of a natural resource may rise in response to changing world market conditions. In the static case, the unemployment rate would drop as wages rise, in some cases sharply, such as in response to booming commodity prices (e.g., oil, gold). In a dynamic labor market, however, the local labor force itself is not fixed. Instead, word about new employment opportunities will spread, and the local labor supply will expand. We can in fact expect three distinct adjustments or labor responses: First, previously discouraged workers return to the labor market, thus contributing to a larger LF (e.g., at LF' in Fig. 3.5). Second, more workers may find it worthwhile to commute into the region from a different labor market. And, third, new migrants may arrive in the local labor market from elsewhere in the nation. How these relative numbers get sorted out has important implications for how the benefits of the new employment opportunities are distributed among local residents of the labor market and others. Note that this question about distributional effects would not be relevant at the national level.

So far, we have assumed that individual firms are price takers in the local labor markets within which they operate. In other words, they pay the same amount per unit of labor as other firms, and their individual hiring and firing decisions have no influence on local wages. While this is certainly a reasonable assumption for national markets (with possible exceptions of highly skilled occupation such as basketball players), an analysis of local labor markets opens up interesting new possibilities and questions. In particular, it is plausible that a single large employer such as a big-box

Fig. 3.5 Equilibrium wage and employment determination

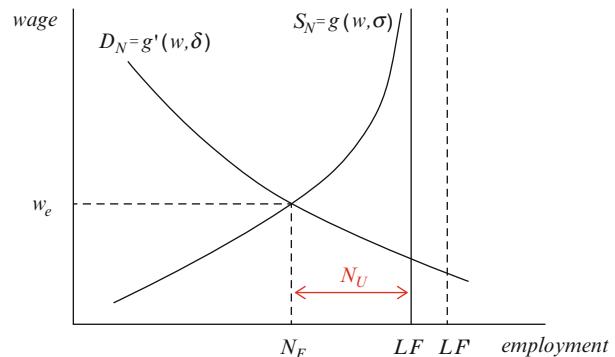
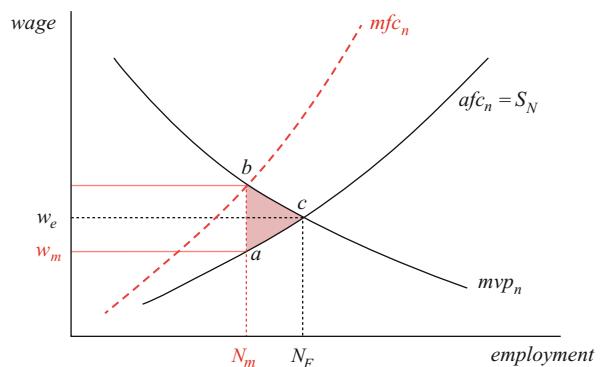


Fig. 3.6 Labor monopsony showing deadweight loss to society



retailer or mine operator is the only major local employer, who in that case does influence wages with its hiring decisions. In particular, starting from the initial competitive position, this employer would have to start paying higher wages to attract workers. In practice, this means that the firm faces an upward-sloping average labor (or factor) cost curve, rather than a horizontal line at the market wage, which would represent a perfectly elastic labor supply to the firm (i.e., the perfectly competitive case). The average labor supply curve (afc_n) can only slope upward if the marginal labor supply (mfc_n) curve lies above it, as is shown in Fig. 3.6.

The profit-maximizing monopsonist chooses to hire labor up to the point where the marginal factor cost is equal to the marginal value product; in other words, at the optimum the cost of the last unit of labor is equal to the contribution to profits of that unit. In Fig. 3.6, this occurs at level of employment N_m , which is clearly below the competitive level of N_F . At this point, it is also clear that $w_m < w_e$, so that the monopsonist both pays less and hires fewer workers than the firm facing competition from rivals. The deadweight loss to society from this outcome is given by triangle abc in Fig. 3.6. This area shows the additional gain to society from hiring more workers and expanding output ($\int_{\Delta N} mvp_n$ where $\Delta N = N_F - N_m$) compared to the additional cost to society of employing these labor resources ($\int_{\Delta N} S_N$) which would prevail under perfect competition.

The existence of monopsony situations may not be far-fetched if we are studying local labor markets (Edwards 2007). For example, transportation costs may pose insurmountable barriers in certain remote regions, including those within Appalachia in the USA, thereby limiting the labor response or supply elasticity. Also, Bonanno and Lopez (2012) found that Wal-Mart's power to set prices in labor markets may be greater in rural areas and in the Southern USA, where the company has operated for more years than anywhere in the world. Another barrier may arise in the form of entry costs into a market in which a particular firm dominates the hiring activity, thus limiting competition. This could relate to a labor skill that is so specialized as to not be useful in other fields. We will return to this in Sect. 3.4 in the context of examining the benefits of agglomeration.

3.3 Defining Labor Market Areas

In practice, the concept of local or regional labor market areas is difficult to define precisely. Even nations that on the surface have clearly delineated labor markets experience flows of migrant workers across their borders, which raises the question of how far labor market areas extend. On the other hand, workers in so-called integrated labor market areas such as the EU also face barriers including those related to languages and potential cultural differences. Thus, even at the national level, it is not always obvious where a labor market area (LMA) begins and ends.

Within nations, the answers are not much clearer. Researchers have used administrative boundaries ranging from multistate regions to states, statistical metropolitan areas, and individual counties to implicitly or explicitly define LMAs, for example. Yet labor markets also exist within larger cities within single counties, and they can extend internationally. Others function only at certain times of the year – for example, those for seasonal workers such as migrant farm labor – and there are specific definitions of labor market areas for certain occupations that do not exist for others. For example, lawyers need to pass bar exams in the state(s) in which they wish to practice, and medical doctors can receive special benefits by locating in so-called areas of physician shortages, usually in rural regions.

In any regional analysis that seeks to group “basic spatial units” into meaningful regions, such as LMAs, it is important to distinguish between homogeneity versus functionality of the regions (Hoover and Giarratani 1984, Chap. 9). The underlying rationale is that “regions” should consist of spatial units that are more homogenous within than across regions. An example of a homogenous region is the former US manufacturing belt or the wheat-growing area of the Northern Great Plains. These regions would be affected in similar ways by policy changes, such as interest rate policies that influence the foreign exchange value of the dollar or new labor market policies that alter unionization laws. Alternatively, regions may exhibit higher or lower degrees of integration with respect to functions such as cross-border commuting, trade, and telecommunication flows. For example, Metropolitan Statistical Areas are characterized by spatial subunits (counties) that are linked by commuting or newspaper circulation patterns, that is, economic functions; usually, these

involve a central node in the form of a business district and outlying or peripheral bedroom communities. At the end of Sect. 3.3.2, the implications of distinguishing between functional and homogenous regions for the analytical methods used are presented.

3.3.1 Historical Efforts

US government agencies and researchers have sought to define regional labor markets as far back as the 1940s. The War Manpower Commission “defined a labor market as the widest area within which employees with fixed addresses would accept employment,” while the War Labor Board “defined a labor market area as one in which the wage structures and levels in an industry were fairly uniform” (Minnesota University Employment Stabilization Research Institute 1948, p.1; this section draws heavily on Goetz 1999). A related definition holds that workers are part of the same labor market area if they can change their jobs without moving their residence. Later, Smart (1974, 255) pointed out:

[i]t is perhaps surprising that systematic criteria for defining labour market areas have not been more extensively developed. The main obstacle has probably lain in the fluid and heterogeneous character of work movements. The jobs to which workers travel at any moment do not necessarily represent their optimum (achievable) preferences, or those of their employers in recruiting labour, particularly if there are imperfections in the labour market resulting from deficiencies in information. Labour economists ... have shown that the operation of market forces is often extremely imperfect, as seen, for example, in the ranges of earnings which may be found for similar occupations in the same area.

Other variables that have been used to delineate labor markets conceptually include the extent of competition faced from other firms. Horan and Tolbert (1984: 10) define LMAs as “geographic areas within which transactions between buyers and sellers of labor are situated and occur on a regular basis,” or “the area bounded by the commuting radius around a district of concentrated employment opportunities.” Thus, at the core of any geographic boundary to an LMA are the notions of place of residence and place of work and the friction or transaction costs of moving between them. Geographic or economic distance – in terms of travel time and costs of gathering market information – is important.

Commuting is central to defining LMAs because it is inherently spatial, involving the physical separation of place of work and of residence, and it also relates to matching labor supply and demand. Along with population size, commuting is essential in the definition of Metropolitan Statistical Areas (MSAs), where counties are considered part of an urban (metro) core if a certain share of their residents works in the core. Klaassen and Drewe (1973, 21) proposed that regional and local labor markets can be distinguished as follows:

The most significant criterion on which to draw a distinction between local and regional labour markets is distance – physical, or even better economic (using travel costs and intervening economic opportunities) and social distance (communication barriers and travel time).

Although many studies have been conducted using various administrative boundaries to delimit labor market areas by default, including state and county borders, the most prominent work uses various aggregations of counties based on commuter flows to arrive at formal LMAs. The Bureau of Economic Analysis (BEA) draws on central place theory to delineate economic areas (EAs) that “represent the relevant regional markets for labor, products, and information. They are mainly determined by labor commuting patterns that delineate local labor markets...” (Johnson and Kort 2004: 68). These authors start with metro- or micropolitan areas that represent the main nodes or centers of economic activity and then use an iterative procedure to sort counties into Component Economic Areas (CEAs) using commuter flows and data on newspaper circulation, in situations where the commuting data are insufficient to arrive at a classification. The final EA has to be (p. 71) “a region of sufficient size to support regional statistical analyses, and each economic area is a labor market that is independent of other labor markets.” Using the 2000 county-to-county commuter flow data, Johnson and Kort (2004) arrive at 344 CEAs, of which 177 are either not large enough (e.g., <50,000 employed residents or <10 counties and <100,000 employed residents) or have too many out-commuters (>8 %) to qualify as an EA or both. These are subsequently merged with CEAs that qualify as economic areas to arrive at a total of 179 BEA EAs, with maximum inter-EA commuting of $\leq 4\%$. This number is up from 172 EAs that were obtained using the 1990 commuter flows. As an example from Europe, Casado-Diaz (2000) uses travel-to-work data from the 1991 census in Spain to delineate local labor market areas, including those that are specific to different industries and occupations.

3.3.2 Cluster-Based Analysis

A different approach to identifying LMAs is taken in Tolbert and Sizer (1996; for updates see the URL under Fig. 3.7), who argue that the BEA’s focus on large urban centers and its hinterlands inadequately captures labor markets in rural areas. They also start with the commuter flow data from the census and create a frequency matrix capturing the flows between two counties. Entry a_{ij} is the number of commuters from i into j divided by the smaller of the two resident labor forces of the counties; by convention, elements of the main diagonal are set to 0. The authors use PROC CLUSTER within SAS, which uses dissimilarity between elements rather than similarity. For this reason, Tolbert and Sizer (p. 12) transform their data using $d_{ij} = d_{ji} = (1-a_{ij})$, which represent distance coefficients. Thus, the smaller the value of d_{ij} , the closer the distance and the greater the similarity between two counties in terms of the commuting relationship. In this procedure, the goal is for counties that belong to a commuting zone to be more similar than those that do not belong; in other words, the aim is for homogeneity within groups and heterogeneity across groups.

The hierarchical clustering algorithm starts by placing each unit (in this case county) into its own category and then merges counties starting with the pair



Fig. 3.7 Map of US labor market areas (Source: <http://www.ers.usda.gov/briefing/rurality/lmacz/> Accessed 3 January 2012)

exhibiting the greatest similarity or commuting strength. This procedure continues iteratively until all counties have been merged into a single cluster. The history of the merging procedure is recorded in a dendrogram, which resembles a tree and shows the average distance between clusters on the vertical axis and records the different counties on the horizontal axis. The length or depth of the branches of the tree measures the strength of the connection between two or more counties. The only remaining question then is where to choose the cutoff point that determines whether or not a set of counties is in the final cluster. If two or more clusters are sufficiently different or distinct from one another, then they are separated into two different groups.

From this analysis, Tolbert and Sizer generate 741 distinct commuting zones for the USA, based on the 1990 commuting data, compared with 709 using the 2000 data. These CZs are subsequently aggregated into 394 LMAs using the population threshold criterion of $\geq 100,000$ persons, which is substantially more than the number obtained by the BEA and in part reflects the greater sensitivity to inclusion of rural areas. Figure 3.7 shows the map of LMAs developed by Tolbert and Sizer for the USA. Readers familiar with the US county geography will see that these areas appear very different from the typical county-level maps. Indeed, it is difficult to even pick out the state outlines from this figure, because most LMAs straddle state lines. As an artifact of the population settlement history and natural geography of the county, the geographically relatively large LMAs in the western part stand out compared to those in the eastern portion.

It is important to note a key limitation of this analysis, namely, that it does not allow the overlapping of counties across LMAs or cross-commuting across the areas. In other words, the hierarchical procedure allocates counties in such a manner that counties cannot belong to multiple LMAs simultaneously, and the method does not accommodate cross-commuting behavior. Isserman et al. (1986: 544) point out that “overlapping regionalization schemes make modeling far more complex because of the need to assure consistency in the treatment of the areas included in more than one region.” Recent work, which is beyond the scope of the overview that can be presented here, draws on advances in the science of networks to address some of these issues by accommodating both hierarchy and overlap of counties within LMAs. This new work also aims to address issues that are arising in the context of megacities and megaregions, including how they are most appropriately delineated.

The LMAs described here are based on a functional rather than a homogenous regionalization, and it must be noted that the clustering procedures described are not necessarily the best techniques available. Fischer et al. (1993) instead apply the intramax and iterative proportional fitting procedures (IPFP) to telecommunications data in Austria, arguing that these methods have the advantage of not requiring contiguity among spatial units. Using these methods, the authors find (p. 225) “...important differences in terms of both the data transformation and the grouping techniques used” and suggest that these methods may be better suited to modeling flows of data, commuters, or goods and services across spatial lines, that is, for functional delimitations of regions. These methods employ more refined procedures for creating standardized network matrices (than, e.g., those used in Tolbert and Sizer 1996) that are then used in grouping the spatial units into regions. The intramax procedure compares expected and observed flows to gauge the interactions among spatial units, while the IPFP involves iterative convergence in the matrix standardization to an equilibrium criterion; this procedure is available in SAS as PROC IPFPHC.

In another paper exploring less restrictive clustering procedures, Baumann et al. (1983) describe a method for functional regionalization that (p. 54) “neither assumes symmetry in the daily commuting relationship between the basic spatial units nor requires a priori assumptions about the relative importance of each basic spatial unit.” The contribution of this paper lies in showing that how regions are defined and identified in labor market analyses can affect both the parameters of estimated regression models and how the models perform statistically. More specifically, analysts need to choose (p. 58) both the optimal number k of labor market regions, in the first step, and how best to aggregate these regions in the second step; these are known as the scale and aggregation problems, respectively. Interestingly, because there also is no contiguity constraint in the IPFP approach adopted, “...the result [in terms of LMA identified] may sometimes be a functional regional typology and not a functional regionalization” (p. 61). These authors conclude that the defining of regions is complementary to econometric estimation, and both procedures should be considered jointly. This is an area warranting additional research.

Finally, Barkley et al. (1995) examine how much spatial association in the form of spread and backwash effects exists within labor market areas that include various definitions of Functional Economic Areas (FEAs). The authors argue that “[i]f there is little revealed dependence or association [within an economic area], then the integrated functional region concept is of limited value in the analysis of hinterland development problems” (Barkley et al. 1995: 298). Presumably, if the association is high, on the other hand, then the LMAs are useful analytical tools. Barkley et al. (1995) calculate the local and global Moran’s *I* and the Getis-Ord *G* statistic for population and income changes within eight FEAs to arrive at two central conclusions. First, they find that linkages between the core and periphery varied in a statistically significant manner both across and within their chosen LMAs. Second, spatial economic linkages to the urban core area tended to favor those hinterland areas that were closest to the core. Thus, with current definitions, relationships within FEAs (and thus LMAs) tend to be robust for counties closer to the urban core and less so for more remote rural counties. In turn, this suggests that spatial socioeconomic analyses should take into consideration counties’ position in the spatial rural–urban hierarchy, in addition to considering FEAs.

3.4 Labor Market Area Analyses

Having outlined theoretical concepts related to spatial or regional labor markets and methods for defining them empirically, we are now ready to review applications in the form of models that have been developed by researchers. The first section examines models used to explain differences in labor earnings across LMAs, while the second considers economic models of boom and bust.

3.4.1 Explaining Differences in Labor Earnings

One of the active areas of research that seeks to explain earning differences across LMAs builds on Mincer (1974) – style earning equations of the following sort:

$$\ln w_i = \alpha_0 + \alpha_1 \text{edu}_i + \alpha_2 \text{exp}_i + \beta \mathbf{X}_i + \varepsilon_i \quad (3.6)$$

where w_i is a measure of average earnings in labor market area i , edu_i is average education (usually in years) of the population, exp_i is years of on-the-job or similar experience, X_i is a vector of other individual characteristics that affect earnings, and ε_i is a random error term. While some recent studies are conducted at the level of individuals, others have been conducted at the level of labor market areas, such as the US states or rural versus urban areas, including cities (e.g., Goetz and Rupasingha 2004; Moretti 2010). Studies have found, for example, that returns to education in rural labor markets are lower than commensurate returns in urban areas. In particular, a one percentage point increase in educated adults in rural areas leads to only about a one-third of the comparable increase in earnings of more urban areas.

While this literature describes the issue, it does not fully explain why this happens; partial explanations focus on job access elsewhere by, for example, statistically interacting the educational attainment variable with interstate highway access, which reduces travel cost to a distant job, possibly in an urban core area. In some ways, this literature is a precursor to studies that focus on understanding how agglomeration economies can lead to higher labor productivity and earnings in more densely settled places. This subject is discussed next.

The fact that labor (and firm) productivity is higher in more densely settled areas, or labor markets, has been established going at least as far back as Marshall (1890). Such agglomeration economies result from better sharing, matching, and learning that occur when businesses are located near one another (Rosenthal and Strange 2004; Puga 2010). Although these three forces may apply to inputs generally, they are especially powerful in the case of labor. For example, greater intensity of economic activity allows for greater specialization in production as well as in learning highly specialized skills, and the knowledge that there are numerous sources of demand for a particular skill makes it easier for a worker to justify the cost and risk associated with investing in such a skill. Note also in terms of the earlier discussion related to monopsony or oligopsony labor hiring power of firms that this means a lower likelihood of firms extracting rents from workers. Puga (2010: 213) further notes that “concentrations of employment iron out idiosyncratic shocks and facilitate the transfer of labor from low to high productivity establishments.” The so-called “thick” labor markets, finally, also facilitate searches and improve matches between employers and workers, making mismatches in terms of skill requirements less likely.

3.4.2 Models of Spatial Adjustment: Booms and Busts

One area of research in which local labor market modeling is especially important is in so-called boom and bust studies, which often revolve around natural resource-based exogenous price shocks but which are also relevant to understanding the local economic effects of government investments and industrial recruitment. Current oil and natural gas drilling activity in the Marcellus (e.g., Pennsylvania) and Bakken (e.g., North Dakota) Shales, for example, has brought about profound local economic adjustments that are best studied in the context of labor markets. In North Dakota, the attraction of new population via in-migration has been so phenomenal that the state’s population recorded an all-time high in the 2010 census, after experiencing decades of out-migration. Earlier booms and busts during the 1970s energy crisis related to coal in Appalachia have also been studied using labor market models.

Often these kinds of studies are interested in finding out how a boom in one sector (e.g., natural resources or manufacturing, when a plant is recruited from elsewhere) impacts other sectors within the same local labor market area. For example, if a state provides subsidies to recruit a manufacturer, existing firms may see their local labor costs bid up as the demand for labor increases, especially if there are no labor supply adjustments – for example, through in-migration or

changes in labor force participation. Likewise, in an economic bust, the impact on local wages will depend on the degree to which the local labor market adjusts to the decline, for example, through out-migration or early retirement of workers from the labor force.

A general model used to study the effect of a commodity- or government investment-related boom and bust on key local labor market variables is the following (Black et al. 2005):

$$\Delta \ln(Y_{ist}) = \sum_{j=1}^3 \beta_j(T_i P_{jt}) + (\text{State}_s \text{Year}_t) \zeta + \varepsilon_{ist} \quad (3.7)$$

where $\Delta \ln(Y_{ist}) = \ln(Y_{ist}) - \ln(Y_{ist-1})$ is the first difference in the dependent variable, Y_{ist} , which may be jobs, wages, or total earnings; i indexes a county, s a state, and t is the year; β_j and ζ are coefficients to be estimated, where $j=\{1,2,3\}$ for a boom, peak, and bust, respectively (the peak occurs between the boom and the bust period); $T_i = 1$ when county i is in the treatment group and 0 otherwise; $P_1 = 1$ during the boom period and zero otherwise and analogously for the peak and bust; State_s is a vector of the states included in the analysis; Year_t is a vector of the years covering the boom-peak-bust period; and ε_{ist} is an error term indexed for each county, state, and year.

The product $\text{State}_s \text{Year}_t$ holds constant those factors that vary at the state level over time (a state-time fixed effect). Here, the county is the basic labor market area (unit), and it is nested within the larger state context, with different policies and institutions. Note also that this setup incorporates a counterfactual comparison of treated with nontreatment counties and that the coefficients β_1 , β_2 , and β_3 “measure the difference in average growth between the treatment and comparison counties during the boom, peak and bust, respectively” (Black et al. 2005: 457).

The analysis can be extended to examine spillover effects from the basic (usually tradeables) to the nonbasic (local goods) sectors within each local labor market area (i.e., county). This is done using the same equation as above but including earnings and employment changes for the non-mining sector only as the dependent variables, and it allows a comparison of growth in these variables in the treated versus non-treated counties to assess any spillovers. In addition, the non-mining sector can be further decomposed into separate sectors such as services, retailing, construction, and manufacturing to study impacts in greater detail. Lastly, the above analysis can be adapted to examine the effects of booms and busts on wages, poverty rates and levels, and population migration (see Black et al. 2005 for details).

A number of important local labor market concepts, including commuting, unemployment, labor force participation, and migration, can be jointly studied at the county level using the following identity which is due to Partridge et al. (2009):

$$N_F = (N_F/N_R)(N_R/LF)(LF/POP)POP \quad (3.8)$$

This equation is an identity as appropriate canceling of terms will reveal, and with the exception of N_R , each of these variables has been defined previously.

The term N_R measures employment by place (county) of residence as opposed to place of work. In the USA, the labor force is counted both by the Bureau of Labor Statistics at the place or county of residence and by the Bureau of Economic Analysis at the place or county of employment, through firm-level surveys. This makes it possible to create a crude *commuting* measure in that $N_F/N_R > 1$ characterizes a situation of average net in-commuting into the county, while a reversal of the inequality corresponds to net out-commuting.

Next, the term N_R/LF captures the employment rate by place or county of residence of the worker, where both of these are measured, or one minus the *unemployment rate*, $(1 - v) = (1 - N_U/LF)$. The *labor force participation* rate, or $\varphi = LF/POP$, was described earlier, while changes in population (*POP*) will occur in response to *migration*. This can be seen clearly once we log-differentiate the above equation, which allows us to study the effect of adjusting each of these four key components on employment at the firm (note also that $\Delta \ln X \approx \Delta X/X$). This gives us the rate of change in each of the component variables, as follows:

$$\begin{aligned}\Delta N_F/N_F &= (\Delta N_F/N_F - \Delta N_R/N_R) + (\Delta N_R/N_R - \Delta LF/LF) \\ &\quad + (\Delta LF/LF - \Delta POP/POP) + \Delta POP/POP\end{aligned}\tag{3.9}$$

which is change in (net commuting + [1–unemployment rate] + labor force participation + net migration), where we have abstracted from births and deaths, which are the other major demographic components that lead to population change. Thus, we have decomposed, for a given local labor market, the key components that lead to adjustment in firm-level employment in response to a particular economic shock. The effects of each of these components can in turn be estimated from county-level data, using the following relationships (Partridge et al. 2009: 15):

Commuting: $(\Delta N_F/N_F - \Delta N_R/N_R) = \beta_{COM} \Delta N_F/N_F$

Employment rate: $(\Delta N_R/N_R - \Delta LF/LF) = \beta_{ER} \Delta N_F/N_F$

Labor force participation: $(\Delta LF/LF - \Delta POP/POP) = \beta_{LFP} \Delta N_F/N_F$

Migration: $\Delta POP/POP = \beta_{MIG} \Delta N_F/N_F$

where each change equation in addition features a specific county fixed effect, a fixed time effect, and a conventional residual error term. Because $\sum \beta = 1$, the estimated regression parameters can be used to calculate the share of new labor supplies coming from changes in each of the four sources identified.

This kind of study also confirms the importance of considering different geographic scales in studying LMAs, and it reveals that longer time lags produce different results because they allow more adjustments to occur between labor markets. Sensitivity to adjustments across LMAs in turn is important for fully assessing the distributional impacts of local and regional economic development projects.

3.5 Conclusions

Even though LMAs are not treated explicitly in standard micro- or macroeconomic textbooks and defining them remains an ongoing challenge, this chapter has

provided a sense of the importance and merits of considering different labor markets over space and their individual characteristics, as well as the adjustments that inevitably occur within and between them over time. Although researchers have developed detailed LMA models and insights into how they operate, more work is needed to classify these markets and to understand how they are affected by federal policy as well as other exogenous shocks in a dynamic setting. With growing apprehension about income inequality in the USA especially, regional labor market analyses promise to generate a better understanding of the causes and consequences of such economic disparities. The questions surrounding local market power, supply adjustments, short- and long-term boom-bust effects, and the causes and consequences of agglomeration economies, especially with respect to labor and learning spillovers, also remain as important research topics. Finally, considerable promise lies in applying emerging insights from network science to better formulate and delineate LMAs and associated megacities and megaregions. This is an area of research that will likely receive much attention in the future.

References

- Barkley DL, Henry MS, Bao S, Brooks KR (1995) How functional are economic areas? Tests for intra-regional spatial association using spatial data analysis. *Papers Reg Sci* 74:297–316
- Baumann JH, Fischer MM, Schubert U (1983) A multiregional labour supply model for Austria: the effects of different regionalisations in multiregional labour market modeling. *Papers Reg Sci Assoc* 52:53–83
- Black D, McKinnish T, Sanders S (2005) The economic impact of the coal boom and bust. *Econ J* 115:449–476
- Bonanno A, Lopez RA (2012) Wal-Mart's monopsony power in metro and non-metro labor markets. *Reg Sci Urban Econ* (in press) doi 10.1016/j.regsciurbeco.2012.02.003
- Cahuc P, Zylberberg A (2004) Labor economics. MIT Press, Boston, MA
- Casado-Diaz JM (2000) Local labour market areas in Spain: a case study. *Reg Stud* 34(9):843–856
- Edwards ME (2007) Regional and urban economics and economic development: theory and methods. Taylor and Francis, New York
- Fischer MM, Essletzbichler J, Gassler H, Trichtl G (1993) Telephone communication patterns in Austria: a comparison of the IPFP based graph-theoretic and the intramax approaches. *Geog Anal* 25(3):224–233
- Glaeser EL (2008) Cities, agglomeration and spatial equilibrium. Oxford University Press, Oxford
- Glaeser EL (2011) Triumph of the city. Penguin Press, New York
- Goetz SJ (1999) Migration and local labor markets. In: Loveridge S (ed) The web-book of regional science. Regional Research Institute, Morgantown. <http://www.wvu.edu/~regional/WebBook/Goetz/contents.htm> Accessed 3 Jan 2012
- Goetz SJ, Rupasingha A (2004) The returns to education in rural areas. *Rev Reg Stud* 34:245–259
- Goetz SJ, Partridge MD, Rickman DS, Majumdar S (2011) Sharing the gains of local economic growth: race to the top versus race to the bottom economic development. *Environ Plan C* 29(3):428–456
- Hoover EM, Giarratani F (1984) An introduction to regional economics, 3rd edn. AA Knopf, New York
- Horan PM, Tolbert CM (1984) The Organization of work in rural and urban labor markets. *Rural Studies Series*. Westview Press, Boulder, CO
- Isserman A, Taylor C, Gerking S, Schubert U (1986) Regional labor market analysis. In: Nijkamp P (ed) *Handbook of regional and urban economics*, vol 1. Elsevier, Amsterdam

- Johnson KP, Kort JR (2004) 2004 Redefinition of the BEA economic areas. *Surv Curr Bus* 84:68–75
- Klaassen LH, Drewe P (1973) Migration policy in Europe. Saxon House/Lexington Books, Farnborough/Lexington
- Marshall A (1890) Principles of economics. Macmillan, London
- Mincer J (1974) Schooling, experience and earnings. Columbia University Press, New York
- Minnesota University Employment Stabilization Research Institute (1948) Local labor market research: a case study: The St. Paul Project, 1940–1942. University of Minnesota Press, Minneapolis
- Moretti E (2010) Local labor markets. NBER WP 15947. <http://www.nber.org/papers/w15947> Accessed 15 Dec 2011
- Partridge MD, Rickman DS, Li H (2009) Who wins from local economic development? A supply decomposition of US county employment growth. *Ec Dev Quart* 23:13–27
- Puga D (2010) The magnitude and causes of agglomeration economies. *J Reg Sci* 50:203–219
- Rosenthal SS, Strange WC (2004) Evidence on the nature and sources of agglomeration economies. In: Henderson JV, Thisse JF (eds) *Handbook of regional and urban economics*, vol 4. Elsevier B V, Amsterdam
- Smart MW (1974) Labour market areas: uses and definition. *Prog Plan* 2(4):238–353
- Tolbert CM, Sizer M (1996) US commuting zones and labor market areas: a 1990 update. ERS Staff Paper 9614, Washington, DC

Alessandra Faggian

Contents

4.1	Introduction	60
4.2	The Standard Job Search Model	61
4.2.1	Basic Formulation	61
4.2.2	Extensions	63
4.3	The Matching Function	65
4.4	Job Search and Migration	69
4.5	Conclusions	72
	References	72

Abstract

This chapter summarizes the main developments in job search theory ever since its inception in the 1970s. After describing the assumptions and formulation of the basic model, the chapter moves onto analyzing how the original framework has been extended by removing some of the initial limitations. A separate section is then devoted to the matching function theory which represents one of the main developments of job search theory in more recent years and whose importance has been recognized by the award of the 2010 Nobel Prize in economics. The last section attempts to reconcile job search and migration theory by introducing the role of space and describing the main contributions on these topics by regional economists.

A. Faggian

AED Economics, Ohio State University, Columbus, OH, USA

e-mail: faggian.1@osu.edu

4.1 Introduction

With two thirds of national incomes coming, on average, from labor, it is not surprising that labor economists have devoted so much effort in modeling job search decisions. Although the share of labor income in more developed countries has decreased in recent years, labor still represents the main means of support of most households. Hence, choosing the “right” job is one of the most important lifetime decisions individuals have to make.

Job search theory became popular in the 1970s as an alternative to the “standard” *neoclassical labor supply theory*. The neoclassical framework, based on the assumption of perfect information, did not allow for unemployment where individuals *actively* sought work but were unable to find it. Individual agents only had two options, either being employed or being inactive (i.e., not part of the labor force). However, evidence showed that unemployment and its duration were not negligible. This led a group of scholars to formulate an alternative theory able to account for unemployment, which became known as “job search theory.” The main premise of job search models is that looking for a job is a dynamic sequential process and that individuals have to decide when to stop this process under conditions of uncertainty and imperfect information. Frictional unemployment is a natural outcome of this process.

Ever since the 1970s, job search theory has been refined and extended in several directions and countless contributions have been published on the topic. While most of these contributions are interesting and provide new insights into the labor search process, the development of a “matching function” stands out as being possibly the most fundamental development in job search theory since its inception. The importance of this latest development was recognized by awarding the 2010 Nobel Prize to Peter Diamond, Dale Mortensen, and Christopher Pissarides, whose work formed the basis of such development.

This chapter tries to survey the key features of the job search theory while facing some constraints. Firstly, the literature on job search theory is so vast that only the milestones can be presented. Secondly, job search models involve a high degree of mathematical sophistication, which goes beyond the scope of this chapter. In presenting the models, the mathematical formulation is streamlined to a minimum while preserving the main insights of each model. The economic meaning and intuition behind the mathematical formulation is also provided.

The organization of the chapter is chronological going from the initial contributions in the 1970s to the more recent contributions (also known as the Diamond-Mortensen-Pissarides model) for which the 2010 Nobel Prize was awarded. The last section before the conclusions will introduce the role of migration into job search theory and will present some contributions developed specifically in the regional economics field.

4.2 The Standard Job Search Model

4.2.1 Basic Formulation

Although the origin of job search theory is normally attributed to the two seminal articles by McCall (1970) and Mortensen (1970), two papers by Stigler (1961, 1962) questioned the perfect information assumption of the neoclassical theory and laid some of the foundations of what was then developed in the 1970s.

Both in Stigler's model and in the basic job search model, the individual has more than one earning opportunities available and has to select the "best" one. However, the "strategy" to select the best job is different. In Stigler's model, the main decision an individual has to make is how many jobs to sample before deciding which one is the "best." Sampling an extra job has an associated "search" marginal cost c over a given time period, and the decision variable is the sample size n representing the number of firms a job seeker will consider in their search. The neoclassical assumption of perfect information is a special case where the cost c equals zero.

In job search models, the decision process is sequential. There is no "optimal" sample size because the jobs are randomly sampled one at a time and the individual stops when an acceptable job becomes available. Hence, the number of jobs sampled depends on their sequence and the sample size itself is a random variable (Mortensen 1986). The basic job search model is simply an "optimal stopping rule" problem which can be described as follows.

Each job seeker receives n wage offers – w_1, w_2, \dots, w_n – per period of length h spent searching for a job. The best offer in each period is equal to

$$w = \max \{w_1, w_2, \dots, w_n\} \quad (4.1)$$

Wages associated with future job offers are distributed according to a probability distribution $f(w)$. The job seeker's aim is to maximize net benefits (future stream of income minus search costs).

In its simplest form, the job search model is based on the following assumptions:

- (i) Time is continuous; t denotes the time periods, each of length h .
- (ii) Although wages associated with future job offers are unknown to the seeker, *the probability distribution $f(w)$ is known* and it is constant over time.
- (iii) The *search cost* per unit of time is c .
- (iv) Once the seeker *accepts* a job offer, this leads to *permanent employment* at a fixed per-period wage, w .
- (v) The discount rate is r .
- (vi) Individuals have infinite lifetimes.
- (vii) The seeker receives one job offer per period.
- (viii) If the job is *rejected*, it *cannot be recalled*.
- (ix) The seeker is unemployed.

Based on these assumptions, the basic model can be formalized as follows.

$W(w)$ is an unspecified functional relationship representing the future stream of income associated with a per-period wage equal to w . The present value of the lifetime wealth, at time t , is

$$V_t = -ch + e^{-rh}W(w) \quad (4.2a)$$

if the individual accepts the job offer at time t or

$$V_t = -ch + e^{-rh}E(V_{t+1}) \quad (4.2b)$$

if the individual rejects the job offer at time t and continues the search at time $t+1$.

The best strategy is one which maximizes V_t or, in other words,

$$V_t = -ch + e^{-rh} E\{\max[W(w), V_{t+1}]\} \quad (4.3)$$

where $E\{.\}$ denotes the expectation operator. Remember that, by assumption, the cost of search, c , is constant over time, the wage distribution is constant over time, and the individual has an infinite lifetime. We know that

$$V_t = V_{t+1} = V \quad (4.4)$$

And so it follows that

$$V = -ch + e^{-rh}E\{\max(W(w), V)\} = -ch + e^{-rh}E\{\max(W(w) - V, 0)\} \quad (4.5)$$

In the continuous time version of the model with infinite time horizon, this formula simplifies to

$$rV = -c + E\left\{\max\left(\frac{w}{r}\right) - V, 0\right\} \quad (4.6)$$

An individual would accept the job offer if $w/r > V$ and reject it if $w/r < V$. The value which separates acceptable and unacceptable job offers, $w^* = r/V$, is called reservation wage.

It can be shown that, in the simplest case in which we assume the individual has just one job offer per period, the reservation wage equals

$$w^* = -c + \frac{1}{r} \int_{w^*}^{\infty} (w - w^*)f(w)dw \quad (4.7)$$

The economic interpretation of Eq. (4.7) is intuitive. It says that the marginal cost of continuing searching, given by the opportunity cost w^* and the cost of search c , is equal to the expected marginal return from continuing the search process, which is given by the second term on the right-hand side of Eq. (4.7).

4.2.2 Extensions

Violation of Assumption (v): Finite Lifetime (Gronau 1971)

One of the assumptions of the basic job search model, presented in the previous section, is that individuals have infinite lifetime. The violation of this condition is analyzed in Gronau (1971). His contribution shows that if individuals have a finite number of periods, the reservation wage does not level out to an equilibrium level but rather decreases over time as we approach the end of the period.

Violation of Assumption (vi): The Seeker Receives a Different Number of Offers in Each Period

The best way of modeling the variation in the arrival rate of offers is to assume that they follow a Poisson distribution. Following Mortensen (1986), we can define the probability distribution, $q(n)$, of the number of job offers received, n , in each period of time as a Poisson distribution with parameter λ :

$$q(n) = e^{-\lambda} \frac{\lambda^n}{n!} \quad (4.8)$$

And Eq. (4.7) will simply become

$$w^* = -c + \frac{\lambda}{r} \int_{w^*}^{\infty} (w - w^*) f(w) dw \quad (4.9)$$

where λ is the offer arrival rate which is inversely related to the expected length of time between job offers.

Violation of Assumption (vii): Allowing Recall

In the basic model where most of the parameters are unchanging over time, allowing recall does not make any difference. A job offer, which is rejected today, will be rejected also in the future, as the parameters upon which the individuals are basing their decision are the same. However, in some more sophisticated versions of the model, recall might make a difference in the reservation wage expression.

Violation of Assumption (viii): On-the-Job Search (Burdett 1978)

One of the most criticized assumptions of the basic search model is that only unemployed individuals search for jobs. In his contribution Burdett develops a model which removes this constraint by allowing workers to look for another job while employed. In his model, employed workers have to decide their job search “intensity” by allocating their time to three alternative uses: working, enjoying leisure, and seeking another job. The search intensity decision is modeled by allowing the costs of search, c , and the arrival rate of offers, λ , to be both a function of search intensity, s . Hence, the reservation wage expression is modified as follows:

$$w^* = -c(s) + \frac{\lambda(s)}{r} \int_{w^*}^{\infty} (w - w^*) f(w) dw \quad (4.10)$$

Both the cost of search and the number of job offers received are increasing in s . The optimal search intensity declines with the wage earned which also implies that an employed individual has a lower optimal search intensity than somebody who is unemployed. If the wage earned is high enough, the employed worker will not search ($s = 0$). In other words, on-the-job search is always associated with a “bad” job position. The Burdett model offers an alternative explanation to why more experienced workers have generally higher salaries. The traditional wisdom – based on the human capital theory – is that productivity increases with experience. While that is generally true, the Burdett model also highlights that older workers – having been in the labor market for longer – are also more likely to have found a better job offer. The Burdett model also assumes implicitly that some nonwage characteristics of the job become known to the worker only after having started working. Moreover, as Blau (1992) points out, spending some time on the job might be beneficial to remove the *stigma* associated with being unemployed. A final point is that while quits are not allowed in the basic standard job search model, now they are a possibility (*violation of assumption (iii)*). Although we do not enter into the details of his work, Wilde (1979) provides a good discussion of job quits and nonwage job characteristics.

Violation of Assumption (i): Unknown Wage Distribution (Rothschild 1974)

The most fundamental assumption of the standard job search models is that the distribution of wages is known to the seekers and constant over time. Many have criticized this assumption. High (1983) points out that relaxing this assumption makes the job search models more realistic. The first contribution to study the problem of unknown distributions is Rothschild (1974). Although his example was related to prices and not wages, the same principles apply to wage distributions. With unknown wage distributions an optimal single reservation wage does not exist. This is because seekers learn about the distribution as they progress with their search and hence they change their reservation wage based on the new information collected. Rothschild formalized it in the following way. Suppose search is without recall, eliciting each time an element (price or wage in our case) of the finite set:

$$P = \{p_1, p_2, \dots, p_n\} \quad (4.11)$$

Suppose also that the number of times each price (or wage) has been observed is summarized by

$$N = \{N_1, N_2, \dots, N_n\} \quad (4.12)$$

Let us define $\rho = \frac{1}{\sum_{i=1}^n N_i}$ and $\mu_i = \rho N_i$ with $i = 1, \dots, n$ so that (μ, ρ) contains all the information accumulated by an individual where μ is the content of the information and ρ is its precision. $\lambda(\mu, \rho)$ is a probability distribution which represents the seeker’s beliefs based on what he has already observed. Based on this probability distribution, an individual would expect to observe a price (wage)

p_i with odds $\frac{\lambda_i(\mu, \rho)}{1 - \lambda_i(\mu, \rho)}$. Every time a new price (wage) p_i is observed, the new information is assimilated into the model according to the following function:

$$h_i(\mu, \rho) = \left(\frac{\mu_1}{\rho + 1} + \dots + \frac{\mu_n}{\rho + 1}; \frac{\rho}{\rho + 1} \right) \quad (4.13)$$

The optimal strategy is defined – as normal in this kind of dynamic models – by induction. Applying the Rothschild model to wages rather than prices, we can define the expected wage if the seeker is only allowed to search once as

$$V_1(\mu, \rho) = \sum_{i=1}^n \lambda_i(\mu, \rho) w_i \quad (4.14)$$

and the expected wage if the individual is allowed to search T times as

$$V_T(\mu, \rho) = \sum_{i=1}^n \lambda_i(\mu, \rho) \max\{w_i, V_{T-1}[h_i(\mu, \rho)] - c\} \quad (4.15)$$

The value of the extra information converges over time $V_T(\mu, \rho) \leq V_{T-1}(\mu, \rho)$ so that

$$\lim_{T \rightarrow \infty} V_T(\mu, \rho) = V(\mu, \rho) \quad (4.16)$$

The optimal stopping rule is based on a comparison between the actual wage offer received and the expected wage offer minus the costs of search. An individual has to stop searching when a wage offer satisfying this condition is received:

$$w_i \geq V_{T-1}[h_i(\mu, \rho)] - c \quad (4.17)$$

Empirical papers using unknown wage distributions often rely on experimental methods to test the behavior of individuals. One good example is Cox and Oaxaca (2000), which also provides a review of previous experimental empirical contributions on this issue.

4.3 The Matching Function

The early job search models are one-sided models where the main problem is to find an optimal stopping rule for individuals looking for a job. In this sense, these models are supply-side oriented. The work by Peter Diamond (1982a, b), Dale Mortensen (1970, 1986), and Christopher Pissarides (1979, 1984, 2000) – who were awarded the 2010 Nobel Prize for economics – changed this logic allowing job seekers and firms to interact in a framework that became known as “matching theory.”

Matching theory is based on a “matching function” whose simplest form is

$$M = m(U, V) \quad (4.18)$$

where $m(\cdot)$ is an unspecified functional relationship, U is the number of unemployed workers looking for job, V is the number of vacancies advertised by firms, and M is the number of matches between the two sets. The matching function is a kind of a black box and it is assumed to have constant returns to scale so that – on average – an unemployed worker finds a job in a unit period length with probability

$$\frac{m(U, V)}{U} = m(1, \theta) \equiv \alpha(\theta) \quad (4.19)$$

and a vacancy is filled with probability

$$\frac{m(U, V)}{V} = \frac{m(U, V)}{U} \frac{U}{V} \equiv \frac{\alpha(\theta)}{\theta} \quad (4.20)$$

The parameter $\theta = \frac{V}{U}$ is a measure of “labor market tightness.” Every time there is a “match” between a vacancy and a job seeker, a surplus is formed (relative to the situation in which they remained “unmatched”). This surplus is split through a Nash bargaining process. The Nash bargaining process is a very well-known concept in game theory and it refers to the work of John Nash (1953) for which he received the Nobel Prize in economics in 1994. The Nash bargaining process is a way to formalize the interaction of individuals using a simple two-player game. Each player contracts with the other on the basis of their preferences as described by their utility functions. A good description of the concept can be found Osborne and Rubinstein (1994). A high value of θ means that it is easy for a job seeker to find a new job. This means that the job seekers are in a relatively strong position compared to the firms offering the jobs and can therefore negotiate to get a higher share of the surplus generated by the match.

In its simplest version, the model assumes that both firms and workers are homogenous, each firm consists of a single job, and there is a unit measure of workers. Workers move from unemployment to employment at an endogenous rate $\alpha(\theta)$ and in the opposite direction at an exogenous rate λ . The expected lifetime utilities, given a discount rate r , of being unemployed and employed – U and N respectively – are

$$rU = z + \alpha(\theta)(N - U) \quad (4.21)$$

where z is an instantaneous return, e.g., from home production, and

$$rN = w + \lambda(U - N) \quad (4.22)$$

Similarly, the firm hires at an endogenous rate $\alpha(\theta)/\theta$ and loses workers at an exogenous rate λ . The expected profits associated with the vacancy being not filled and filled – V and J , respectively – are

$$rV = -c + \frac{\alpha(\theta)}{\theta}(J - V) \quad (4.23)$$

where c is the cost incurred while the vacancy is unfilled, and

$$rJ = (y - w) + \lambda(V - J) \quad (4.24)$$

and $(y - w)$ is the surplus generated to the firm with income y while the worker is employed at a wage w . In equilibrium $V = 0$, so after solving for V in Eq. (4.24) and substituting into Eq. (4.23), the following can be derived:

$$\frac{\alpha(\theta)}{\theta} = \frac{r + \lambda}{y - w} c \quad (4.25)$$

Equation (4.25) describes the relationship between “market tightness” and wage. When the wage is higher, a firm is less willing to post vacancies (lower θ). However, to find the steady-state values of (θ, w) , we need a second condition to be combined with Eq. (4.25). This comes from the observation that the surplus to be divided between the worker and the firm when a match is realized is the worker surplus plus the firm surplus:

$$\frac{w - rU}{r + \lambda} + \frac{y - w}{r + \lambda} = \frac{y - rU}{r + \lambda} \quad (4.26)$$

Assuming the worker gets a β proportion of this surplus, we arrive at a second relationship between θ and w

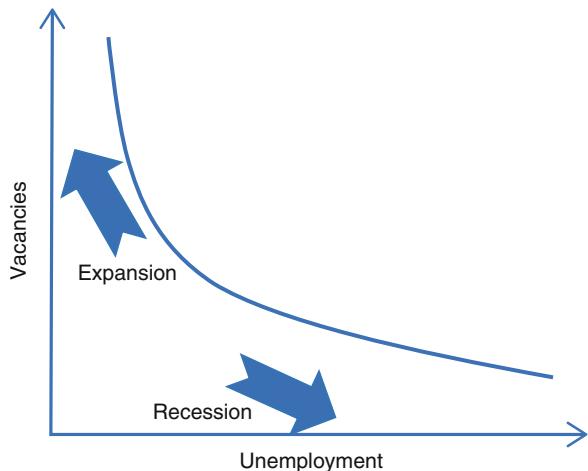
$$w = \beta y + (1 - \beta) \frac{(r + \lambda)z + \alpha(\theta)w}{r + \lambda + \alpha(\theta)} \quad (4.27)$$

Solving the system given by Eqs. (4.23) and (4.25) gives us the steady-state conditions for θ and w .

This basic matching model, which has also been called by the Diamond-Mortensen-Pissarides (or DMP) model has been extended in several ways. Two very detailed reviews of the matching function approach and its extensions are given by Petrongolo and Pissarides (2001) and Albrecht (2011). Petrongolo and Pissarides (2001) in particular review how the empirical contributions have tried to operationalize the matching function model. They classify the empirical studies into four categories:

1. Contributions which rely on the estimation on a Beveridge curve
2. Contributions which estimate aggregate matching functions
3. Contributions which estimate matching functions for local labor markets
4. Contributions which focus on estimating hazard functions for unemployed workers

The Beveridge curve is the equilibrium relationship between vacancy rate and unemployment rate. In a perfectly functioning (neoclassical) labor market, there would be no unemployed individuals looking for jobs nor vacancies to be filled,

Fig. 4.1 Beveridge curve

but – as pointed out by the job search theory – frictions are a reality and the Beveridge curve is a representation of this. Estimated Beveridge curves slope downward and we observe movements along the curve toward the right in a recession (where fewer vacancies are posted and more individuals are unemployed) and toward the left when an economy is booming (Fig. 4.1).

The Beveridge curve (Blanchard and Diamond 1989) is consistent with the expectations of the matching theory and hence provides indirect support for it. However, the majority of studies seeking support for the DMP model rely on the direct estimation of aggregate matching functions. As for production functions, a functional form needs to be chosen for matching functions. The majority of studies use a log-linear specification for the matching function with constant returns to scale, although some other specifications have also been tested. Petrongolo and Pissarides (2001) list 17 empirical studies which try to estimate aggregate matching functions, and they conclude by stating that plausible values for the unemployment elasticity range from 0.5 to 0.7.

Comparing the estimates for local labor market matching functions is considerably more complicated as the studies vary greatly in terms of countries studied and the level of aggregation. In their extensive review, Petrongolo and Pissarides (2001) also list 11 sectoral matching function studies covering a range of countries including the UK, the USA, the Netherlands, the Czech Republic, and Slovakia. They conclude that, although there is a great variety in the periods analyzed, the level of aggregation used, and the variables used in the matching functions, these studies do not contradict the main findings of the more aggregate studies.

Finally, contributions on individual hazard rates are microeconomic in nature and stress mainly the effect of individual characteristics on matching probabilities. One of the advantages of these studies is that they try to distinguish between the variables influencing the probability of *receiving* a job offer and those influencing the probability of *accepting* a job offer once received. Age, education, and

experience have all been proved to greatly influence the likelihood of receiving offers, while the cost of search, the unemployment income, and the overall perceived distribution of wages affect the individual reservation wage and hence the willingness to accept a job offer.

4.4 Job Search and Migration

For regional economists, there is one element missing in job search theory and that is *location*. Migration might improve the chances of finding a good match or might be the result of having found a good match elsewhere. In applying the job search theory to migration, a fundamental distinction needs to be introduced: whether the migration is part of the search (ex ante) or a consequence of the job search process (ex post). Molho (1986), in his review, labels these two as “speculative” and “contracted” types of migration. Speculative migration is “*undertaken in the hope of finding a suitable opportunity at the point of destination*” and contracted migration is “*undertaken after having already secured such an opportunity*” (Molho 1986 p. 402).

Several authors considered the case of speculative migration. Rogerson (1982), for instance, included in his job search model spatial variables such as the distribution of wages for each region and a matrix of costs related to distances. McCall and McCall (1987) presented a sophisticated model in which they also assumed that migration must precede job search. Their model is based on a “multiarmed bandit methodology” combined with the theory of belated information. The multiarmed bandit methodology takes its name from a traditional slot machine (one-armed bandit), and it is used in statistics to describe sequential optimization processes. Multiarmed bandit models are usually applied to problems where an agent is trying to simultaneously acquire new knowledge and maximize his/her utility based on the existing knowledge already acquired. Some of their findings are that regions with large wage variability attract migrants, while regions with large nonpecuniary returns are more likely to have high rates of both in- and out-migration.

Other authors tackled the problem of contracted migration. Gordon and Vickerman (1982), for instance, focus specifically on this issue and build a general model in which the probability that an individual z will migrate from origin a to destination b , at time t , is decomposed as follows:

$$P_t(m_{ab}^z) = P_t(s_a^z)P_t(o_b^z|s_a^z)P_t(d_b^z|o_b^z)P_t(1 - R_a^z|d_b^z) \quad (4.28)$$

$P_t(m_{ab}^z)$ is the probability that individual z will migrate ($m = 1$ if migrating; zero otherwise) from location a to a location b at time t ; $P_t(s_b^z)$ is the probability that individual z located in a is searching ($s = 1$ if searching; zero otherwise) for a job at time t ; $P_t(o_b^z|s_a^z)$ is the probability that individual z receives a job opportunity ($o = 1$ if receiving an opportunity; zero otherwise) in b conditional to the probability of being in search; $P_t(d_b^z|o_b^z)$ is the probability that the z individual accepts ($d = 1$ if the individual accepts; zero otherwise) the opportunity that was given; and finally

$P_t(R_a^z)$ is the probability that individual z residing in a has already received an acceptable opportunity at time t ($R = 1$ if the individual already has an acceptable opportunity; zero otherwise). These probability functions are not defined explicitly in Gordon and Vickerman (1982).

More recently Basker (2003) tried to reconcile speculative and contracted migration by proposing a model in which the question of whether migration is part of the job search or a consequence is endogenously determined by the model itself. In Basker's model an individual z , looking for a job, has three possible options: (a) search locally, (b) search globally (i.e., also outside the region of domicile) from home, and (c) search globally by moving preemptively. The optimal strategy is based on the maximization of individual utility. Each of the three search options has an associated utility function – U_L , U_G , and U_M , respectively – which depends upon several parameters: the value of the reservation wage, the relative favorableness of the local market, the probability of finding a job globally, the costs of search, and the cost of migrating (for a full specification of the utility functions and the model setup, see Basker 2003, pp. 4–5). Local search is assumed to be costless, but is less likely to yield a job (especially a high-paying one). If the local market conditions are really bad, all but the lowest skilled choose the last strategy: migrating to search for a job elsewhere. As the favorableness of the local labor market increases, the probability of speculative migration decreases and more high-skilled workers turn toward contracted migration (i.e., a global search without migrating first). Eventually, very favorable local labor markets mean that everyone adopts the “search locally” strategy.

Jackman and Savouri (1992) is an interesting contribution that combines the matching function approach with migration. Migration is defined in their model as a way to solve a spatial mismatch. In Jackman and Savouri's words, migration is a “*special case of job-matching in which a job-seeker in region a is matched to a job in region b* ” (p. 1434) and therefore depends on the number of unemployed people (U) in the origin region a and the job vacancies (V) in the destination region b . This is different from the traditional matching function approach where a match does not necessarily imply a geographical movement. A match can indeed happen in the area where the worker lives and/or has been working previously. The concept is formalized by defining a generic “hiring function” – whose functional form is not specified – as in Eq. (4.29):

$$H_{ab} = H(U_a, V_b) \quad (4.29)$$

with $\partial H_{ab}/\partial U_a > 0$ and $\partial H_{ab}/\partial V_b > 0$.

The model is quite innovative because it considers regional level variables (at UK Government Office Regions level, which is comparable to the European NUTS1 level) rather than individual characteristics, but the motivation for the need of such a model is rather weak. Jackman and Savouri (1992), indeed, begin by stressing that their model represents an alternative to the traditional *human capital migration model* (Sjaastad 1962), which fails to explain the direction of interregional flows in a recession. According to the human capital migration theory,

migration can be seen as an investment in the human agent, which has costs and renders returns. A person will decide to migrate when the net present value of a migration investment is positive. Let us suppose that a potential migrant wants to move from region a to region b . He/she will migrate only if the net present value (NPV) of his expected returns in region b (destination) is greater than that in region a (origin) minus the cost associated with relocation (C_{ab}), i.e., $NPV_b > NPV_a - C_{ab}$.

Jackman and Savouri argue that, since regional differences are highest in a recession, the human capital model forecasts that more people would move from poorer to richer regions, but the evidence shows that actual migration flows tend to “rise in times of prosperity and fall in a recession” (p. 1433) when they are most needed to restore balance to the system. The hiring function approach provides an explanation to these perverse migration flows by assuming that the number of engagements falls in a recession.

Although the authors are correct in pointing out that the human capital migration model, as such, is inadequate to explain the actual patterns of migration flows observed during recession, the human capital framework can easily be adapted to fit these facts. The human capital approach is neoclassical in essence so the decision to migrate depends exclusively on the comparison of future net benefits associated with the decision to move (Sjaastad 1962). The probability of finding a job is set equal to one and is unaffected by macroeconomic conditions. The addition of a probability function depending on the status of the economy could solve the inadequateness of the human capital migration model to explain lower migration flows in a recession without the need of a completely new alternative model. Especially in the case where the person is actually employed in the region of origin (enjoying a certain future income stream even though it may be low), there is less incentive to move because the probability of finding a job elsewhere is lower.

The reasons for the negative relationship between increased regional gaps in recession and lower probability of finding a job by migrating includes the Jackman and Savouri (1992) argument that employers react to crisis by reducing recruitment and therefore jobs become more difficult to find. As a result, the role of information costs also needs to be considered. Information costs increase when jobs become more sparsely distributed throughout the territory. Moreover, since people perceive that there is a crisis, their reservation wages normally go down. Jobs with lower wages may be more easily available locally and this in turn reduces the chances of having to make a migratory move. Properly defining the function for probability could then reconcile the job search and human capital theories of migration.

Despite the fact that the human capital and job search theories are often regarded as competing, they reach similar conclusions regarding migration. First of all, they both predict that individuals with higher human capital are more likely to migrate. In the case of the human capital theory, this is due to the fact that individuals have to be compensated for their investment in education, and in the case of job search, they need to be compensated for their higher reservation wage. However, one difference needs to be emphasized. In the human capital theory, the migration propensity of each single individual increases with education, while in the traditional job search

theory, on average, higher-human-capital individuals are more mobile than lower-human-capital individuals, but this does not necessarily hold true for every single individual. Indeed, whether or not an individual migrates is related to the location of the *first* acceptable job (i.e., the job that meets the reservation wage). Jobs are randomly distributed over space and the process is sequential (one offer at a time), so it *may* be that some individuals are lucky enough to find an acceptable offer close to their current location. However, higher-reservation-wage jobs are expected to be more sparsely distributed in space so that, on average, higher-human-capital (and therefore higher-reservation-wage) individuals have to move further.

4.5 Conclusions

The aim of this chapter was to present the main ideas behind job search theory and its importance in the field of economics. Job search theory, though microeconomic in nature, contributed to explain macroeconomic phenomena such as frictional unemployment, which could not be explained by the traditional neoclassical theory. Since its inception, there have been many extensions to the model. For example, on the theoretical front, the heterogeneity of individuals has been emphasized, while other contributions focused on “family” job search, in which the decision regarding a job is not taken by an individual but rather by the whole household. On the empirical side, the availability of better data – both individual and aggregate – provided the basis to test some of the propositions of the models. In recent years, many empirical contributions employed experimental methods to better understand individual behavior in the labor market. While this chapter only scratched the surface of job search theory, it hopefully provided the basic notions for further study.

Acknowledgments I acknowledge the support of research grant ECO2010-16006 by the Spanish Ministry of Science.

References

- Albrecht J (2011) Search theory: the 2010 Nobel memorial prize in economic sciences. *Scand J Econ* 113(2):237–259
- Basker E (2003) Education, job search and migration. University of Missouri working paper, 02–16
- Blanchard OJ, Diamond PA (1989) The Beveridge curve. *Brook Pap Econ Activity* 1:1–76
- Blau DM (1992) An empirical analysis of employed and unemployed job search behavior. *Ind Labor Relat Rev* 45(4):738–752
- Burdett K (1978) A theory of employee job search and quit rates. *Am Econ Rev* 68(1):212–220
- Cox JC, Oaxaca RL (2000) Good news and bad news: search from unknown wage offer distributions. *Exp Econ* 2(3):197–225
- Diamond PA (1982a) Aggregate demand management in search equilibrium. *J Polit Econ* 90(3):881–894
- Diamond PA (1982b) Wage determination and efficiency in search equilibrium. *Rev Econ Stat* 49(2):217–227

- Gordon I, Vickerman R (1982) Opportunity, preference and constraint: an approach to the analysis of metropolitan migration. *Urban Stud* 19(3):247–261
- Gronau R (1971) Information and frictional unemployment. *Am Econ Rev* 61(3):290–301
- High J (1983) Knowledge, maximizing, and conjecture: a critical analysis of search theory. *J Post Keynesian Econ* 6(2):252–264
- Jackman R, Savouri S (1992) Regional migration in Britain: an analysis of gross flows using nhs central register data. *Econ J* 102(415):1433–1450
- McCall JJ (1970) Economics of information and job search. *Q J Econ* 84(1):113–126
- McCall BP, McCall JJ (1987) A sequential study of migration and job search. *J Labor Econ* 5(4):452–476
- Molho I (1986) Theories of migration: a review. *Scott J Polit Econ* 33(4):396–419
- Mortensen D (1970) Job search, the duration of unemployment, and the Phillips curve. *Am Econ Rev* 60(5):847–862
- Mortensen D (1986) Job search and labor market analysis. In: Ashenfelter O, Layard R (eds) *Handbook of labor economics*. North Holland, Amsterdam, pp 849–920
- Nash J (1953) Two-person cooperative games. *Econometric Society, Econometrica* 21(1): 128–140
- Osborne MJ, Rubinstein A (1994) *A course in game theory*. MIT Press, Cambridge, MA
- Petrongolo B, Pissarides C (2001) Looking into the black box: a survey of the matching function. *J Econ Lit* 39(2):390–431
- Pissarides C (1979) Job matchings with state employment agencies and random search. *Econ J* 89(356):818–833
- Pissarides C (1984) Search intensity, job advertising and efficiency. *J Labor Econ* 2(1):128–143
- Pissarides C (2000) *Equilibrium unemployment theory*, 2nd edn. MIT Press, Cambridge, MA
- Rogerson P (1982) Spatial models of Search. *Geogr Anal* 14(3):217–228
- Rothschild M (1974) Searching for the lowest price when the distribution of prices is unknown. *J Polit Econ* 82(4):689–711
- Sjaastad L (1962) The costs and returns of human migration. *J Polit Econ* 70(5):80–93
- Stigler GJ (1961) The economics of information. *J Polit Econ* 69(3):213–225
- Stigler GJ (1962) Information in the labor market. *J Polit Econ* 70(5):94–105
- Wilde LL (1979) An information-theoretic approach to job quits. In: Lippman SA, McCall JJ (eds) *Studies in the economics of search*. North-Holland, Amsterdam, pp 35–52

Jan Rouwendal

Contents

5.1	Introduction	76
5.2	The Monocentric Model	76
5.3	“Wasteful” Commuting	78
5.4	Transport Modes, Sorting, and Urban Sprawl	83
5.5	Density, Diversity, and Agglomeration	85
5.6	Owning, Renting, and Unemployment	88
5.7	Conclusions	90
	Appendix Computation of the Reservation Wage	90
	References	90

Abstract

In the monocentric model, commuting is viewed as a burden whose cost shapes the spatial structure of cities to a considerable extent. This view has been challenged by the finding that actual commuting patterns are far from efficient. However, this “wasteful” commuting is better interpreted as an indication of labor market frictions that are traded off against commuting frictions than as a neglect of commuting costs. Urban sprawl results from the decreasing importance of physical space that was the consequence of the automobile and is fundamentally consistent with the basic insights of the monocentric model. Large and diversified urban labor markets flourish when space restrictions are relaxed because this facilitates the matching of jobs and workers along other dimensions. Having a large mortgage puts more stress on this allocation mechanism.

J. Rouwendal

Department of Spatial Economics, VU University, Amsterdam, The Netherlands
e-mail: j.rouwendal@vu.nl; jrouwendal@feweb.vu.nl

5.1 Introduction

Before the industrial revolution, most people lived where they worked. New production techniques and – later – increasing welfare resulted in the spatial separation of the residential and work locations. Commutes provide the connection between the housing and labor markets. The home-work trip is generally considered as a burden: spatial separation causes friction that can only be overcome by accepting transport costs. This view on commuting has long been dominant in the thinking of urban economists. However, commuting also offers some flexibility: one can change job while staying in the same house and vice versa. In dense urban areas, all kinds of jobs are available at reasonable commutes. This second view on commuting has become more prominent in the more recent literature. It does not necessarily contradict the first one: a worker may dislike commuting while at the same time appreciating the job opportunities that a large metropolitan labor market offers. In this chapter, both points of view will be discussed.

In the next section we start with a discussion of the monocentric model that gives a central role to commuting costs in its explanation of the spatial structure of cities. [Section 5.3](#) provides a discussion of the challenge that the discovery of “wasteful” commuting implied for the established view and the answer provided by search theory. [Section 5.4](#) deals with the decrease in commuting cost that was associated with the automobile which had, in accordance with the main insights of the monocentric model, an enormous impact on spatial urban structure. In [Sect. 5.5](#), the advantages of density for matching heterogeneous workers and jobs are discussed, as well as the interaction between agglomeration effects – which tend to favor monofunctional areas – and commuting disutility, which tends to favor mixed land use. [Section 5.5](#) continues with a review of the discussion about Oswald’s thesis which says that homeownership has negative effects on labor market performance and argues that the distinction between outright and leveraged owners is of crucial importance here. [Section 5.6](#) concludes.

5.2 The Monocentric Model

The monocentric model, developed by Alonso, Muth, and Mills, studies the housing market that emerges around an employment center. It thus investigates the connection between a very simple labor market – where identical workers are employed at the same location and earn the same wage – and a somewhat more elaborate housing market, where houses differ in quality as well as in the distance to the employment center. Housing requires land, and land is available in limited quantity around the employment center. Workers prefer to reside close to their work location, but the limited supply of land makes it impossible for all to realize this desire. Commuting provides the possibility to separate the residential and work locations, but workers dislike it. Equilibrium therefore requires that workers with a long commute are somehow compensated. This works via the housing market: cheap housing

compensates for long commutes and allows all households to reach the same utility level, even though their circumstances are quite different.

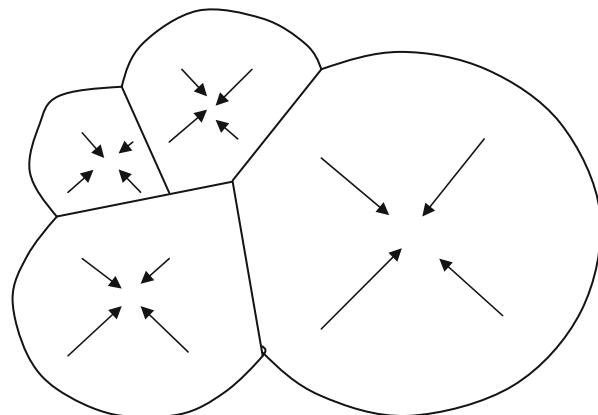
It is useful to go a little bit into the formalities of the model since this clarifies the fundamental role of commuting cost in the model. Households have preferences over housing h and other consumption c that can be summarized in a utility function $u(c, h)$. The budget constraint says that total expenditure on housing and commuting must be equal to net income, defined as the difference between the wage w and the commuting cost. The latter is equal to the product of the distance x to the employment center and commuting cost per unit of distance t . The unit price for housing p is allowed to vary with distance to the center. The budget constraint can thus be written as $c + p(x)h = w - tx$. Note that the unit price for other consumption has been normalized to 1. Maximization of utility subject to the budget constraint leads to the indirect utility function that gives the maximum utility the consumer is able to reach at given location, housing price, and net income: $u = v(w - tx, p(x))$. The equilibrium condition requires indirect utility to be independent of the location; hence, $\partial v / \partial x = 0$. This implies the well-known Muth condition

$$\frac{\partial p}{\partial x} = -\frac{t}{h(p(x), \bar{u})} \quad (5.1)$$

The left-hand side is the slope of the house price function, while the right-hand side equals minus the ratio of the transport cost per unit of distance and the Hicksian demand function for housing. \bar{u} denotes the equilibrium value of utility. The Muth condition determines how the house price changes with distance to the employment center. The house price determines, jointly with the equilibrium level of utility, the demand for housing at each location in the city. If one introduces housing construction in the model, then the housing price determines the density of housing – the size of houses and of gardens close to the city edge but also building height close to the employment center – as well as the population density and the price of land at all locations. In short, virtually every aspect of the housing market in the monocentric city is determined to a considerable extent by the Muth condition in which the commuting cost per unit occupies a central place. It is therefore no exaggeration to say that commuting is a crucial element in this model which can entirely be regarded as focusing on the relationship between the housing and labor market in a simplified setting.

In the monocentric model the edge of the city is determined by the point where the bid rent – the highest rent that is compatible with reaching the equilibrium level of utility – for residential land equals the price of agricultural land. In a city with homogeneous workers, the bid rent equals the housing price function, whose slope is determined by the Muth condition. It is not difficult to show that the size of the city – measured by the distance from the employment center to the edge – increases when transport costs decrease. This simple comparative static result can be related to the phenomenon of urban sprawl, and it demonstrates once again that the Muth condition, with its central role for commuting cost, does a very good job in predicting empirical regularities.

Fig. 5.1 Figure shows four employment centers of different sizes surrounded by disjoint recruitment areas. Wages differ between the centers. In each recruitment area the logic of the monocentric model holds. The arrows indicate the direction of the commutes



Although the model is called monocentric and has often been criticized for being so, it is useful to note that it has no difficulty with the existence of multiple employment centers. If we stick to the assumption of identical workers and allow them to be mobile between the various centers, the logic of the model implies that there will be residential areas around every center from where workers commute exclusively to that center, as is illustrated in Fig. 5.1. The equilibrium condition that all workers must reach the same utility level determines the maximum bids of workers in a particular center for each possible residential location. The worker that offers the highest bid “wins” the residential location, and in this way, recruitment areas emerge for each employment center.

5.3 “Wasteful” Commuting

The central position of the commuting cost in the monocentric model makes clear why Hamilton's (1982) finding, that actual commuting patterns appear to be closer to those resulting from random matching than to an efficient allocation of workers to jobs, was a cause of real concern among urban economists. Hamilton developed a variant of the monocentric model in which some employment is located outside the center. The labor market is competitive, and the wage of the decentralized jobs must therefore be such that any worker who is employed there must be indifferent between this job and one in the CBD. This is a strong condition: it implies that workers accept only decentralized jobs that are located on the straight line between their home and the CBD. All commutes are therefore in the direction of the CBD, although not all of them end there. Armed with this result, Hamilton could show that this commuting pattern is efficient in the sense that it minimizes the total commuting distance traveled by the workers in the city.

This prediction of the extended model could be tested by comparing the actual total commuting distance with the minimum. It was of course not a complete surprise that the actual situation differed from the efficient one. However, the difference was so large that Hamilton's results called into question the logic

of the monocentric model: crisscross commuting seemed incompatible with a strong impact of transportation costs on urban structure. Briefly, most commuting appeared to be wasteful instead of efficient.

A few years later, White (1988) pointed out that Hamilton had ignored the existence of multiple employment centers with their own recruitment areas. The decentralized employment in Hamilton's model differs from the multicentered city with several employment centers, each with their own recruitment area that we briefly discussed at the end of the previous section. In such a city, Hamilton's equilibrium condition does not hold: commutes are not all directed to one particular center. White (1988) presented alternative figures about commuting across boundaries of zones within urban areas that suggested that actual commuting patterns were much closer to the efficient lower bound than was implied by Hamilton's results (see also Hamilton 1989). Again a few years later, Small and Song (1992) confirmed this finding but also showed that within zones there was a considerable amount of "excess" commuting, a term preferred by these authors to Hamilton's adjective. This normatively neutral qualification was probably related to the observation that the assumptions underlying Hamilton's conclusions were quite strong. Real world urban labor markets are characterized more by substantial heterogeneity of both jobs and workers than by the extreme homogeneity that is the standard assumption in the monocentric model. Housing is also an extremely heterogeneous commodity. In addition, it is durable and transformations are costly. These properties make it unlikely that actual urban labor markets will be able to come close to the efficiency boundary by swapping jobs or houses between workers whenever that leads to shorter commutes to both parties involved. The presence of two worker households adds to these problems.

However, these considerations do not answer the question whether the actual commuting patterns could be reconciled with the logic of the monocentric model that attributes a large role to commuting costs. It turns out that a somewhat different view on the labor market can do just this. In the course of the 1970s, search models became quite popular in (nonspatial) labor economics. These models stressed the information problems that occur when workers looking for a(nother) job have to find employers with a suitable vacancy and vice versa. See ► Chap. 4, "Job Search Theory" for a discussion of spatial application of search theory. A spatial version of the standard model of job search developed in this literature runs as follows. Consider an unemployed worker located at x , who is searching for a job. We do not model his or her search activities explicitly, but only the result: job offers arrive now and then. Formally, we assume that there is a constant arrival rate λ that equals the expected number of job offers per period. The jobs offered are identical, except for the net income associated with them. Such differences in net income may result from an identical wage paid in a number of employment centers located at different distances from the searcher's residential location. Alternatively, one may think of a city with decentralized employment à la Hamilton where jobs are located either in the center or elsewhere in the city and the offered wage depends on the location of the job. To keep the model reasonably general, we assume a given distribution of net wages y that will not be specified further at this moment. A job offer is a random

draw from this distribution. Each time a job offer arrives, the searcher has to decide if she accepts it. Acceptance implies the end of the search process and employment in the job for the remainder of her life. Refusal implies that the search process continues, which usually implies the possibility of a better job offer in the (near) future. As long as the search process continues, the worker receives an unemployment benefit b . This benefit must also be interpreted as a net amount of money, as it seems likely that the searcher has to travel in order to locate vacancies. We take the housing consumption h of the searcher as given and assume that there is no saving or income from other sources. That means that instantaneous utility equals $u(b - p(x)h, h)$ as long as the searcher remains unemployed and $u(y - p(x)h, h)$ when a job with net wage y is accepted.

It seems likely that the optimal search strategy is such that if a job with net wage y will be accepted, also all jobs with a higher net wage will be accepted. Unless all job offers are accepted, this must imply the existence of a lowest acceptable wage. Moreover, as long as the arrival rate, the job offer distribution, and the unemployment rate remain unchanged, there seems to be no reason why this critical net wage should change over time. Formal analysis confirms these conjectures and proves that the solution of the dynamic optimization problem has the so-called reservation wage property: there is a critical net wage y^{res} , and the searcher accepts the first offer that implies at least this net wage.

This standard model is consistent with a potentially large amount of apparently wasteful commuting. To see this, consider the situation in which job offers originate from a finite number of unemployment centers indexed $n=1, \dots, N$. (We will not discuss the case in which employment is distributed continuously over space as is, for instance, the case in Hamilton's monocentric model with decentralized employment. The implications are entirely similar). Let φ_n be the probability that a wage offer originates from center n and w_n the wage offered there. The following expression holds: $u(y^{res} - p(x)h, h) = \frac{\rho u(b - p(x)h, h) + \lambda \sum_{n \in A} \varphi_n u(y_n - t d_{xn} - p(x)h, h)}{\rho + \lambda \sum_{n \in A} \varphi_n}$

In this equation, ρ denotes the rate of discount and A the optimal acceptance set, that is, the set of employment centers for which the net wage exceeds the reservation wage. That is, $n \in A$ if $y_n - t d_{xn} > y^{res}$. The distance d_{xn} is the distance between the searcher's residential location and employment center n .

The left-hand side of the equation gives the utility reached by the searcher when the net wage equals the reservation wage y^{res} . The right-hand side shows that it equals a weighted average of the instantaneous utility that is experienced as long as the searcher has not accepted a job offer and the instantaneous utilities that may be experienced after acceptance of a job offer. The appendix provides a simple algorithm for the computation of the reservation wage. It implies that the searcher will only accept job offers from the most attractive employment centers. If the acceptance set has only one element, we are in a situation with exclusive recruitment areas for each employment center. It is not difficult to verify that such a situation will be reached when $\lambda \rightarrow \infty$. Then the searcher can be selective without having to suffer an increase in unemployment duration, and she will accept only job offers that imply the highest possible net wage. When $\lambda \rightarrow 0$, the opportunity cost

of rejecting a job offer becomes very large, and the searcher will accept all job offers that imply a net wage that exceeds the unemployment benefit. The arrival rate thus indicates how far (or how close) the labor market is from the frictionless situation assumed in the conventional monocentric model.

It is interesting to consider the special situation in which identical wages are offered in each center. The reservation wage strategy is then equivalent to a strategy that uses a critical commuting distance: the first offered job that has an implied commute shorter than the critical value will be accepted. In the more general case, when offered wages can vary over centers, long commutes can be compensated by higher wages. Note that the number of acceptable centers can be large, and in special cases with low arrival rates of job offers, the whole set of employment centers in an urban area may be acceptable. If that happens, the worker can commute to any employment center, and his or her behavior is in fact determined by random matching. It is possible that all workers in the urban area are in this position, and in this case the commuting flows are indeed determined by random matching of workers to jobs, the situation that was close to Hamilton's empirical findings.

Does this imply that the connection between the housing market and the labor market that plays such an important role in the monocentric model breaks down? Not at all. To see why this is the case, consider the expression $\sum_{n \in A} \varphi_n u(y_n - t d_{xn} - p(x)h, h)$ which is a weighted sum of the instantaneous utilities that will be experienced after acceptance of a job. It depends on the implied commuting distances from x to the acceptable employment centers and on the price of housing at x . If the searcher would move to a residential location that has better accessibility to employment centers, his lifetime utility would increase, even if the set of acceptable centers would remain unchanged, unless the price of housing would go up. This effect would be reinforced if the searcher becomes more choosy after the movement and changes her acceptance set. Clearly, accessibility to the relevant employment centers is relevant in the present model. Note that this is also the case in the situation in which job offers from all employment centers are accepted (see Rouwendal 1998).

This shows that in a city in which new inhabitants enter the city unemployed and have to choose a residential location, they will prefer those that provide good accessibility to employment centers. Since space is limited, the usual bidding process will then lead to a spatial structure with a trade-off between job accessibility and house prices that is in many respects similar to that of the traditional model. All the conventional predictions about the rent gradient, the density of housing, etc., still follow, even though the workers can now be cross commuting over large parts of or even the entire urban area. One interesting difference is that locations in between employment centers may have reasonably good accessibility to all the centers, without being very close to any of them. The search model predicts that such locations can be very desirable, whereas the traditional theory regards them as inferior to locations in the close proximity of a single center.

The model assumes that households choose a residential location once and for all. Since the model assumes that workers stick to a job forever after it has been accepted, it may be argued that workers have an incentive to move as close as

possible to the job they have accepted. However, this incentive is weakened if we introduce the possibility that workers can lose their job and become unemployed again. It can be shown that little changes in the model developed above when we do so. Sticking to a residential location that has once been chosen may be a good strategy when there are substantial moving costs. Empirically, there is little impact of a change in employment status on the residential location, which suggests that the assumption of a fixed residential location does little harm.

Empirical research based on the search model has also confirmed that workers attach considerable weight to commuting costs when accepting a job. See Van den Berg and Gorter (1997), Rouwendal (1999), and Van Ommeren, Van den Berg, and Gorter (2000). Interestingly, these studies have also shown that a worker's sensitivity to commuting costs depends on household characteristics. A repeated finding is that women attach a greater weight to commuting distance than men and that this is especially the case when young children are present in the household. Allocation of tasks within the household thus appears to interact with labor supply.

The search model introduced above helps to explain the phenomenon of wasteful commuting, but there is certainly also a role to play for heterogeneity on labor supply and demand. This will be further discussed in Sect. 5.5 below. For the moment it is useful to think of the urban labor market as a set of disjoint segments. Workers from one segment cannot be employed in jobs that belong to another segment, and jobs from one segment cannot be filled by workers from any other segment. Within each segment, a search model like the one presented above is valid. In this model, wasteful commuting results from immobility between sectors as well as from search frictions within sectors. If jobs requiring different types of workers are present in the same firms, employers will in general prefer central locations (in the CBD), but there may still exist differences between the worker types that lead to sorting behavior. The slopes of the bid rent curves are now determined not only by commuting costs and housing consumption but also by the frequency and duration of unemployment. This tends to sort workers with a bad labor market position to locations with relatively bad job accessibility.

This mechanism is quite different from Kain's spatial mismatch hypothesis, which is discussed in detail ► Chap. 6, “[Spatial Mismatch, Poverty, and Vulnerable Populations](#)”, although it results in a similar correlation between job accessibility and labor market position. According to Kain (1968), black workers who were often located in ghetto's close to the CBD were disproportionately disadvantaged by the shift of manufacturing industries to peripheral areas, since housing market discrimination hampered their ability to adjust to the new situation by moving closer to the manufacturing jobs. This may well have been true, but it is clear that in general there is a potentially important endogeneity problem involved. Recent research has tried to isolate the effect of living at a location with bad accessibility to jobs by taking advantage of natural experiments. One example is Andersson et al. (2011) who use mass layoff to compare the subsequent unemployment durations of workers at different locations, finding a significant but relatively small effect. Another is Phillips (2011), who reports about a field experiment in which the transport costs of a random subsample of unemployed persons living at remote

suburbs were lowered while using the other part of the sample as a control group. He finds significantly smaller unemployment durations for the treatment group, although the ultimate impact on the share that found a new job was close to zero. It appears therefore that the effects of living at locations with relatively bad geographical accessibility to employment per se are limited, as is implied by the equilibrium interpretation of location choices. The empirical evidence suggests that the structural unemployment among specific groups, like low-educated blacks in many US cities, is not primarily due to their residential locations that do not offer good accessibility to jobs.

5.4 Transport Modes, Sorting, and Urban Sprawl

If it is true that commuting costs are an important determinant of urban structure and that apparently wasteful commuting does not change the essence of this thesis, then one expects that changes in commuting costs will have important consequences for the interaction between labor and housing markets. In this section, we consider various aspects of this issue.

We start with the relationship between income and location choice. Although income does not appear explicitly in the Muth condition, it is not difficult to see how it changes when income increases. There are two effects. First, housing is a normal good, so its consumption increases with income. This also implies that – all else equal – Hicksian demand for housing will be higher for households with a higher income. Second, travel time is an important determinant of the commuting cost, and its value is closely related to income. The first effect tends to make the bid rent curve flatter, while the second tends to make it steeper. It is well known that, in the monocentric city setting, heterogeneous households sort in such a way that the group with the steepest bid rent curve will locate closest to the center. These are the households with the highest incomes if housing is a luxury good, while they are the households with the lowest incomes if housing is a necessity. Since most studies of housing demand find an income elasticity of the demand for housing that is well below 1, the latter situation seems relevant. This suggests that urban economists are again in trouble: empirically the high incomes were the first to suburbanize, and poor households are overrepresented in the city centers. Wheaton (1977) who was one of the first to point this out suggested that the explanation must be found in the durability of housing, which causes old, low-quality housing to be present in the oldest parts of the cities, which are usually close to the employment centers.

However, a closer look at the way new, faster transport modes are introduced in cities reveals a somewhat different insight that was put forward first in LeRoy and Sonstelie (1983). A good example is the early types of public transport, streetcars, that were introduced in a time when most people had to walk to work, keeping the cities small and dense. The introduction of public transport meant that one could move faster, although only by buying a ticket. The relatively well-paid workers have the highest value of time and therefore also the highest willingness to pay for the new transport mode. For these workers the switch to the new transport mode

meant a decrease in the full cost of commuting. This made their bid rents curves flatter, while those of the workers who still walked to work remained unchanged. The logic of the monocentric model thus suggests that the rich who switched to the new transport mode became the group located at the largest distance from the center. It has been documented by Gin and Sonstelie (1992) that this was exactly what happened in nineteenth -century Philadelphia when the streetcar was introduced. Similar stories could be told for many cities where in the nineteenth and early twentieth century, the construction of public transport – commuter –lines that extended to the borders of the existing cities resulted in residential development in the vicinity that was especially used by relatively well-paid workers. This explanation emphasizes that the introduction of a new transport mode flattens the bid rent curve of those who use it while keeping the bid rent curve of the others unchanged. This differential impact on the bid rent curves is the reason why the impact of the introduction of the faster transport mode differs from what is suggested by an analysis of its general impact in the right-hand side of the Muth condition. Ultimately, when the new transport mode is used by almost everyone, this general impact dominates. However, that can take a long time, and the durability of housing may of course contribute to the prolongation of the initial transitory effects.

This history repeated itself when the car was introduced in the early twentieth century. The car is remarkably faster than any previously existing type of public transport and has the important advantage that it could bring one to practically every place as long as good roads are available. That its introduction has had substantial effects on city structure is therefore completely in line with the analysis of the monocentric model. The analysis of the impact of a fast transport mode discussed above suggests that the rich will move out earliest, and that is exactly what happened. Glaeser and Kahn (2003) argue extensively that the automobile is the most important driver of urban sprawl. They emphasize that the increase in average commuting distances that is associated with this phenomenon does not imply an increase in commuting times and indeed that the latter are often shorter for car drivers than for public transport users. Although physical distance has increased, travel time distances rose much less and may even fall. The automobile relaxes the tight connection between the residential and work locations, but it did not change the fundamental forces emphasized by the monocentric model. Commutes remain the essential connections between home and work they have always been since the physical separation of the two that was a main consequence of the industrial revolution.

Two important points remain. First, the analysis suggests that the rich will return to the central city when all workers can afford a car. When the new transport mode becomes available to all, all bid rents curves are flatter and the same relative positions become relevant again, suggesting that the rich will be closer to employment centers than the poor. This may well be related to the surprising revitalization of many inner cities that took place since the 1980s. The “consumer city” that receives much attention in recent work in urban economics may have much to do with it.

Second, the decrease in transport cost that was associated with the car (and the truck) has had enormous consequences for the consumption of land and the size of cities.

It is not to be expected that this development will also be reversed. Only if real transport costs will increase substantially can we expect significantly denser urban areas. Recent analyses of the land use controls also strongly suggest that such measures keep density close to city centers at artificially low levels, thereby contributing to longer commutes and perhaps less well-functioning labor markets in urban areas.

5.5 Density, Diversity, and Agglomeration

The monocentric model takes the location and size of employment centers as given and concentrates on the location of workers around them. The employment centers may be located at points that have particular natural advantages, but there are also endogenous forces at work. Marshall (1890) famously pointed out that specialization of employment centers in particular types of tasks, related, for instance, to a particular type of manufacturing industry, may result in concentration of workers with the complementary skills around these centers. This may cause an agglomeration effect as workers with the specific skills are attracted to the city because it offers them better opportunities to find a job, whereas firms with the specific tasks are attracted to the city because it offers them a better chance to fill their vacancies.

This line of reasoning may be related to our job search analysis by relaxing the assumption that jobs and workers are completely homogeneous. At the end of Sect. 5.3, a brief discussion of a labor market with a number of different segments was provided. In reality, the various segments are seldom completely disjoint, as was assumed there. The heterogeneity in workers and jobs then adds to the difficulties in labor market allocation that were discussed in that section. If jobs differ in tasks to be performed and the suitability of a worker to fill these tasks, vacancies can only be offered to workers whose skills are reasonably close to those required by a particular job, and the wage offer is also likely to depend on the quality of the match. It follows that the heterogeneity of jobs and workers increases search costs. One possibility to mitigate this effect is to let cities or employment centers specialize in particular tasks and skills. The spatial separation of workers and jobs tends to make the labor market more homogeneous, and this facilitates the functioning of the labor market.

To see how this works, return to the model with disjoint segments of the end of Sect. 5.3 and assume now that each segment is located in a different city. For each segment the search model discussed is relevant. For the specialization to have any effect, the specialized cities must offer better possibilities for finding a suitable match than diversified cities of equal size. The mechanism that generates contacts between workers with vacancies and job seekers can be the same in both cities. It may work, for instance, by randomly drawing a vacancy and a job seeker. The probability that such a randomly generated contact implies a reasonable match between skills and tasks is higher in the specialized city than in the diversified one. The result will be a higher arrival rate of job offers for searchers in the diversified city and shorter durations of unemployment and vacancies.

An impact of density on the generation of contacts between labor supply and demand seems quite plausible. In a larger labor market that is geographically concentrated in a small area, it is easier to get into contact with the other side of the market. However, this conjecture is at variance with conventional labor market models in which the contact-generating function has constant returns to scale; see, for instance, Pissarides (2000). Since in such models only the ratio between the number of job searchers and the number of vacancies is important and the same ratio can result from low and high densities, a specialization effect is absent in such models. This constant return to scale property of the matching function implies a congestion effect that seems more plausible in a homogeneous market than in a diversified one. Although the existing empirical evidence, reviewed in Petrongolo and Pissarides (2001), is generally favorable to constant returns to scale specifications, it should be noted that research has not really been focused on the roles of heterogeneity and density.

The diversity of large urban labor markets should be expected to be especially important for the upper end of the skills and tasks distributions where the market is thin and geographical specialization is difficult to realize. A large and diversified labor market then offers firms as well as workers the opportunity to find a reasonable fit between required and available skills. Teulings and Gautier (2004) and Gautier and Teulings (2009) provide an extensive discussion of a search model with increasing returns to scale in the matching function in which cities have particular advantages in terms of labor market allocation. Interestingly, they argue that it is not just physical density, jobs, or workers per squared kilometer but also the integration of geographical locations into a single labor market area that matters. In Gautier and Teulings (2003), they develop an empirical index, denoted as γ , for this aspect that can be estimated as

$$\gamma_n = \frac{\sum_{m=1}^N (s_{nm} - x_m)^2}{1 - \sum_{m=1}^N x_m^2}, \quad n = 1, \dots, N \quad (5.2)$$

The index refers to a (metropolitan) region in which employment and workers are distributed over N areas. In the equation, s_{nm} is the share of the workers in area n that reside in area m , while x_m is the share of houses in the region as a whole (all N areas together) located in m . A lower value of γ_n indicates that the urban area is denser in the sense of being more connected to the areas in the proximity. The index reaches its minimum value, 0, when $s_{nm} = x_m$ for all m . In that situation, the share of workers in n who live in m is equal to the overall share of workers in the city who live in m , which implies that distance does not play a role in attracting workers to area n . This is exactly the apparently random matching of jobs to workers that Hamilton (1982) interpreted as the opposite of efficient commuting. Clearly, the commuting pattern that is extremely inefficient in the context of a perfectly homogeneous labor market is extremely attractive in the context of a labor market with heterogeneity and search frictions. Intuitively, the reason is that the market is better able to reach a good match between skills and tasks when spatial frictions are absent.

In their empirical work, Gautier and Teulings (2003) find a strong negative correlation between their index γ and the log hourly wage, even after the effect of workers or jobs per square km has been taken into account.

Glaeser (1999) argues that one of the advantages of an urban labor markets is that one can switch from one job to another without having to move house. By accepting some inefficiency in commuting, one saves transaction costs in the housing market. Costa and Kahn (2000) have argued that couples of highly educated workers benefit especially from urban labor markets since their often specialized skills make it hard to find a suitable match with available jobs at a reasonable commute in other areas. They show that such couples are indeed strongly overrepresented in large metropolitan areas and argue that this is caused by the dense and diversified demand for labor at these places. Similar concentration effects have been found in other countries, for instance, for the Netherlands. In joint work with Willemijn Weischede, it was shown that commutes of couples are hardly longer than those of single workers in otherwise comparable households, a finding that should probably be attributed to spatial sorting effects. A strategic choice of the residential location within an urban area, which is of course facilitated by their higher income, allows two earner households to keep their commutes limited, their collocation problem notwithstanding.

The discussion thus far has argued that specialization as well as diversity might benefit the functioning of the labor market in cities. There is evidence that both types of agglomeration effects are exploited in reality. For instance, Duranton and Puga (2001) develop a model that explains the empirically documented tendency of many industries to locate in a diversified environment in early stages of their life cycle and in a specialized environment after standardization of their product occurs in a later stage. Although their model does not analyze labor market frictions explicitly, the discussion above fits in the picture they draw. The discussion so far suggests that, at least in diversified cities, a better match between skills and tasks can be realized at the cost of some apparently wasteful commuting. However, a tractable model that deals with the spatial aspects as well as with the match between tasks and skills does not seem to be available.

One should realize that labor market matching is not the only agglomeration force that is active in urban areas. Production externalities tend to have clustering effects on firms, and although the CBD is often treated as a single point in monocentric models, larger concentrations of firms imply in reality longer commutes for their workers. This trade-off is investigated in Lucas and Rossi-Hansberg (2002) who relax the assumption of a given employment center in a setting that is otherwise very close to that of the monocentric model. They attempt to explain the location of workers and jobs in a circular area that is initially completely homogeneous. Two forces are at work. First, workers become more productive when their employment location is close to that of other workers. This agglomeration effect is modeled as a “potential” effect on productivity, and it stimulates clustering of firms. Firms produce a commodity that is sold at the world market without transport costs and at a given price. This means that firms have no preference for locations apart from those that result from the agglomeration forces and the wage they have to pay

to attract workers. Second, workers dislike commuting. This is the force that is also present in the conventional monocentric model. All workers are identical and the labor market is assumed to be perfect. The land market is also perfect and rents are determined by the highest bid.

In this setup it is possible that all firms and workers are spread homogeneously over space. Each worker then lives next door to his job, so there is no commuting. However, firms located close to the border of the city then benefit less from proximity to other workers than those located in the center of the circle. Firms will therefore tend concentrate in the center, but this means that at least some workers have to commute. Concentration of firms implies higher productivity, and therefore higher wages can be offered. But the workers who earn these higher wages can use them to bid more for residential land in the proximity of their employment location, thus counteracting the tendency to agglomeration. The final result of the interaction depends on the relative size of the agglomeration effects and the workers' dislike of commuting. If agglomeration effects are strong, a conventional monocentric city results with all employment concentrated in the center. If commuting costs are more important, other configurations may arise. Lucas and Rossi-Hansberg show, for instance, that it is possible that the city center is a residential area. The inhabitants work in firms located in a ring surrounding the center. These authors also show the possibility of mixed zones in which workers and firms are located next to each other.

The results of Lucas and Rossi-Hansberg (2002) are intriguing since they suggest a rich pattern of spatial equilibria is possible once we relax the assumption of a fixed location and size of employment centers. Although formally their results are restricted to a situation in which space is circular, something similar will probably hold in more general settings. For instance, it is possible that the configuration shown in Fig. 5.1 is consistent with an extended model in which there can also be agricultural land. Recent empirical work, carried out jointly with Hans Koster, confirms the importance of agglomeration as well as dislike of commuting on land prices as determining factors of urban land rents, although other forces, like the presence of consumer amenities and negative externalities imposed by firms on households, should also be taken into account to complete the picture.

5.6 Owning, Renting, and Unemployment

Oswald (1996) has put forward the thesis that there is a causal effect of homeownership on unemployment. According to his analysis, a 10 % increase in the ownership rate increases long-term unemployment by 2 %. Although his paper has long remained unpublished, it was soon referenced in an influential article by Nickell and Layard (1999), and the thesis received a lot of interest. A possible background of a causal effect is the much lower residential mobility of homeowners. If owners would be less willing than renters to accept a job outside their residential area after becoming unemployed, a higher rate of homeownership could indeed push up long-term unemployment rates. Munch et al. (2006) have investigated this hypothesis on the basis of Danish micro data. The theoretical

framework they use is a search model in which job offers can originate from the searcher's region of residence as well as from elsewhere in the country. The first type of job offers can be accepted without moving house, whereas acceptance of the second requires moving to a different region. According to the model, homeowners, who have higher costs of residential mobility, are more reluctant in accepting job offers from other regions than renters. However, their model suggests that they are more willing to accept job offers from the local labor market by setting a lower reservation wage. This effect was confirmed by the empirical analysis, which showed a substantially higher outflow of owners into jobs located in their region of residence. Oswald's thesis was therefore rejected, and this conclusion has been replicated in work for the UK and the Netherlands. These studies find that residential mobility related to accepting a job after a period of unemployment of renters as well as owners is so low that it is questionable whether it can have a significant impact on unemployment rates.

The differences in acceptance of jobs without moving are much more important, and apart from the lower reservation wages that are suggested by the standard search model, they could also be explained by a more intense search effort of owners or differences in the acceptance of long commutes. It is indeed well known that owners have on average longer commutes than renters. However all these explanations fail to make comprehensible the higher overall exit rate from unemployment among homeowners as was pointed out by Van Vuuren (2008). The reason is that the greater effort of owners to find a job on the local labor market is induced by their higher mobility cost, and in economic models such compensating reactions are typically partial. This implies that they predict that overall unemployment spells of homeowners must still be longer than those of renters, whereas the data show the opposite.

Flatau et al. (2003) have found partial confirmation for Oswald's thesis in Australian data: outright owners have longer unemployment spells than otherwise comparable renters. However, they also find that highly leveraged homeowners have shorter unemployment spells and since their number is larger, the net impact of homeownership on unemployment is still negative as in Munch et al. (2006). Recent work, carried out in collaboration with Yuval Kantor and Peter Nijkamp, shows that these findings can be rationalized by a standard search model in which searchers are risk averse if the mortgage payments are larger than the rents and confirmed the strong effects of mortgage payments and also rent subsidies on labor market behavior. In earlier work with Peter Nijkamp, this was shown to be the case for many Dutch households. We also found that highly leveraged homeowners accept long commutes more frequently than others.

The significance of these findings is that they suggest a closer relationship between labor and housing markets than the spatial connection via commutes – however important that is – suggests. In countries with mortgage interest deductibility, many homeowners are highly leveraged, and if this improves their labor market functioning significantly, this should be taken into account when evaluating this measure. At a more general level, Decreuse and Van Ypersele (2011) have recently argued that there is empirically a close connection between housing market regulation and job protection measures.

5.7 Conclusions

This chapter discussed some aspects of the relations between labor and housing markets while focusing on the role of space. Although transportation costs have decreased enormously and information and communication technologies that facilitate cheap and fast interaction between people at different locations have become widely available, the face-to-face interactions with colleagues and clients and suppliers of intermediaries and inputs that often take place at employment locations remain important. Commutes therefore still provide a necessary link between housing and labor market, and there seems to be no reason to expect this to be different in the future although the details of the relationship may change.

This chapter has been limited in the choice of the topics and the literature that is referenced. Although it makes no claim for complete coverage, it hopes to have addressed a number of relevant issues (see Rouwendal and Nijkamp (2004) for an alternative survey). The reader is referred to the other chapters in this *handbook* for discussion of related issues, and the literature that has been cited here contains many references to other papers that help the interested reader to find the relevant literature on topics that could not be addressed here.

Appendix Computation of the Reservation Wage

To see how the equilibrium can be found, start with the simple situation in which there is only one employment center ($N=1$). When the wage offered there is high enough to let the net income at the searcher's residential location exceed the unemployment benefit, the reservation wage will be lower than the net wage offered, and the searcher will always accept a job offer. When there are two or more employment centers, the optimal acceptance set can be determined by the following simple procedure: determine the subset of employment centers whose job offers imply a net wage at x that is at least equal to the unemployment benefit. If it is nonempty, start with an acceptance set that contains only the employment center that offers the highest net wage at x and compute a preliminary reservation wage from the above equation using this acceptance set. Compare this reservation wage with the highest net wage offered by a center that is not yet in the acceptance set. If this wage is higher than the preliminary reservation wage, add this center to the acceptance set. Recompute the reservation wage and repeat this procedure until no employment center fulfills the entrance condition or no employment centers with a net wage at x exceeding the unemployment benefit is left.

References

- Andersson F, Haltiwanger J, Kutzbach M, Pollakowski H, Weinberg D (2011) Job displacement and the duration of joblessness: the role of spatial mismatch. Working paper, US census bureau
Costa D, Kahn ME (2000) Power couples: changes in the locational choice of the college educated, 1940–1990. Quart J Econ 115(4):1287–1315

- Decreuse B, van Ypersele T (2011) Housing market regulation and the demand for job protection. *J Public Econ* 95(11–12):1397–1409
- Duranton G, Puga D (2001) Nursery cities: urban diversity, process innovation and the life- cycle of product. *Am Econ Rev* 91(5):1454–1477
- Flatau P, Forbes M, Hendershott PH, Wood G (2003) Homeownership and unemployment; the roles of leverage and public housing. NBER working paper 10021
- Gautier PA, Teulings CN (2003) An empirical index of labor market density. *Rev Econ Stat* 85(4):901–908
- Gautier PA, Teulings CN (2009) Search and the city. *Reg Sci Urban Econ* 39(3):251–265
- Gin A, Sonstelie J (1992) The streetcar and residential location in nineteenth century Philadelphia. *J Urban Econ* 32(1):92–107
- Glaeser EL (1999) Learning in cities. *J Urban Econ* 46(2):254–277
- Hamilton BW (1982) Wasteful commuting. *J Polit Econ* 90(5):1035–1053
- Hamilton BW (1989) Wasteful commuting again. *J Polit Econ* 67:1497–1504
- Kain JF (1968) Housing segregation, negro unemployment, and metropolitan decentralization. *Quart J Econ* 82(2):175–197
- LeRoy S, Sonstelie J (1983) Paradise lost and regained: transportation, innovation, income and residential location. *J Urban Econ* 13(1):67–89
- Lucas RE, Rossi-Hansberg E (2002) On the internal structure of cities. *Econometrica* 70(4):1445–1476
- Marshall A (1890) Principles of economics. Macmillan, Houndsills
- Munch JR, Roshholm M, Svarer M (2006) Are home owners really more unemployed? *Econ J* 116(514):991–1013
- Nickell SJ, Layard R (1999) Labour market institutions and economic performance. In: Ashenfelter A, Card D (eds) *Handbook of labor economics*, vol III. North Holland, Amsterdam, pp 3030–3084
- Oswald AJ (1996) A conjecture on the explanation for high unemployment in the industrialized nations; Part I. Working paper, University of Warwick
- Petrongolo B, Pissarides C (2001) Looking into the black box: a survey of the matching function. *J Econ Lit* 39(2):390–431
- Phillips D (2011) Getting to work: experimental evidence on job search and transportation costs in Washington, DC. Working paper, Georgetown University
- Pissarides C (2000) Equilibrium unemployment theory. MIT, Cambridge
- Rouwendal J (1998) Search theory, spatial labor markets and commuting. *J Urban Econ* 43(1):1–22
- Rouwendal J (1999) Spatial job search and commuting distances. *Reg Sci Urban Econ* 29(4):491–517
- Rouwendal J, Nijkamp P (2004) Living in two worlds: a review of home-to-work decisions. *Growth Change* 35(3):287–303
- Small KA, Song S (1992) “Wasteful” commuting: a resolution. *J Polit Econ* 100(4):888–898
- Teulings CN, Gautier PA (2004) The right man for the job. *Rev Econ Stud* 71:553–580
- Van den Berg GJ, Gorter C (1997) Job search and commuting time. *J Bus Econ Stat* 15(2):269–281
- Van Ommeren J, van den Berg GJ, Gorter C (2000) Estimating the marginal willingness to pay for commuting. *J Reg Sci* 40(3):541–563
- Van Vuuren A (2008) The relationship between expectations of labor market status, homeownership and the duration unemployment. Working paper, VU University
- Wheaton WC (1977) Income and urban residence: an analysis of consumer demand for location. *Am Econ Rev* 67(4):620–631
- White MJ (1988) Urban commuting journeys are not wasteful. *J Polit Econ* 96(5):1097–1110

Spatial Mismatch, Poverty, and Vulnerable Populations

6

Laurent Gobillon and Harris Selod

Contents

6.1	Introduction	94
6.2	The Theory of Spatial Mismatch	96
6.3	The Empirical Tests of Spatial Mismatch	100
6.4	Local Policies to Reduce Poverty	103
6.5	Conclusions	105
	References	106

Abstract

Spatial mismatch relates the unemployment and poverty of vulnerable population groups to their remoteness from job opportunities. Although the intuition initially applied to African Americans in US inner cities, spatial mismatch has a broader validity beyond the sole US context. In light of a detailed presentation of the mechanisms at work, we present the main results from various empirical tests of the spatial mismatch theory. Since key aspects of that theory remain to be tested, we also discuss methodological approaches and provide guidance for further research. We derive lessons for policy implications and comment on the appropriateness of related urban policies.

L. Gobillon (✉)

Institut National d'Etudes Démographiques (INED), PSE and CEPR, Paris, France
e-mail: laurent.gobillon@ined.fr

H. Selod

The World Bank, PSE-INRA and CEPR, Washington, DC, USA
e-mail: hselod@worldbank.org; hselod@gmail.com

6.1 Introduction

Spatial mismatch is a topic and a theory that relates unemployment and poverty to the structure of cities. It covers a variety of situations according to which the residents of poor neighborhoods are adversely affected by their physical disconnection from places where jobs are located. The focus is thus essentially on large urban areas where such disconnections are likely to be found. Having emerged in the 1960s in the context of racially segregated US cities, the initial intuition quickly became a key topic in urban economics and remained one for more than half a century. Its relevance is now apparent in several other contexts, including for cities of European countries and sprawling metropolitan areas in Asia, Africa, and Latin America.

The spatial mismatch hypothesis was originally formulated by economist John Kain with an initial and exclusive focus on the African American poor in inner cities. The genesis of the hypothesis is rooted in the history of US cities, where, as early as in the 1940s, urban jobs that were initially concentrated in city centers had begun to decentralize to more peripheral locations. This movement went along with the rapid expansion of middle- and upper-class residential suburbs almost exclusively populated by white households. At the same time, the bulk of African Americans were maintaining their residences in city centers, a situation which the author of the spatial mismatch hypothesis attributed to housing market discrimination against blacks that prevented them from suburbanizing to the same extent as whites. The combination of these two trends caused the emergence of the typical US city structure where blacks live far away from the job offers corresponding to their skill levels and that they could apply to. Kain (1968) was the first to hypothesize that the disconnection between places of residence and places of employment could be a key contributor to the high unemployment, low wages, and poverty in the black ghettos of central cities.

A very abundant literature followed Kain's seminal paper for more than four decades and variants were expressed. One noticeable change in focus was the role of race in the "workings" of spatial mismatch. By assuming residential segregation against blacks, the initial spatial mismatch hypothesis clearly put race on the agenda but limited its role to a factor explaining residential immobility. It thus presented race only as a cause of spatial mismatch. After two decades of empirical work, however, whether blacks were really disconnected from or affected by distance to job opportunities became the center of a controversy as a study on Chicago concluded to the opposite and suggested that *race* rather than *space* was in fact the main determinant of the bad labor market outcomes of blacks in inner cities (Ellwood 1986). Following this study, whether spatial mismatch was a relevant explanation of black labor market outcomes polarized the debate for several years in spite of an increasing number of sources documenting the physical disconnection of blacks from jobs (and likely from job opportunities) and although subsequent empirical papers, including on Chicago, were finding that spatial mismatch did play a key role in black unemployment. The opposition between the race and space arguments then gradually disappeared from the literature.

Most contributions to the literature on spatial mismatch are empirical papers that try to assess a link between the disconnection from jobs and bad labor market outcomes (see Ihlanfeldt and Sjoquist 1998 and Gobillon et al. 2007 for extensive surveys). In this literature, the main challenge throughout has been to establish causality and to isolate the contribution of spatial mismatch to labor market outcomes from other spatial and nonspatial explanatory factors. Although some authors have looked at the effect on wages and labor market participation, most papers focus on unemployment so that it is probably not exaggerated to present spatial mismatch as mainly a spatial theory of unemployment. Surprisingly, however – and this is probably one of the few examples in the history of economic theory – it is only starting in the late 1990s, this is to say *after* the publication of many empirical papers on the topic, that the theoretical works on spatial mismatch began to emerge. The publication of spatial mismatch models gave the initial hypothesis the status of a fully fledged theory rather than being just an intuition. These models typically shed light on (a) the *causes* of spatial mismatch, i.e., on why blacks in US cities live in areas that are physically distant from jobs – and in some cases proposing alternative explanations to housing market discrimination – and on (b) the *consequences* of spatial mismatch, shedding light on several competing mechanisms to explain how physical disconnection from jobs can affect the labor market outcomes of black workers. These models provided an analytical framework to think about spatial mismatch. By formalizing the diversity of potential mechanisms, they also provided a sound basis to derive the policy implications associated with the different mechanisms. Models of spatial mismatch also helped clarify several of the drawbacks and misunderstandings regarding the scope and interpretation of related empirical work. These models for instance provided interesting insights on what the counterfactual of spatial mismatch should be and implications for empirical tests: Should one compare the outcomes of black and white subgroups exposed to different levels of disconnection from jobs? Or should the test focus instead on an estimation of what the outcomes of black inner-city unemployment would be under a less intense disconnection from jobs? Theory also helped discard a number of inadequate tests, for instance, the idea that short commutes provide an interpretable indication regarding the level of spatial mismatch (as short commutes may indicate both neighborhood proximity to or remoteness from jobs if the only jobs that remain accessible are the local ones). Spatial mismatch models also paved the way for refined empirical tests of specific spatial mismatch mechanisms.

Over the past decade, new directions in the spatial mismatch literature have also emerged.

Some authors have argued that race and space, rather than being alternative explanations of black unemployment, may *combine* to explain the harmful effects of spatial mismatch. The interaction between race and space may probably reflect several mechanisms, not all of which are clearly spelled out at present. One underlying assumption is that blacks are not affected by distance to jobs in the same way as whites. Another underlying assumption is that proximity to particular types of low-skill jobs may matter. The reason why this should be the case is the subject of recent research and illustrates the tendency of the literature to move

toward the elicitation and exploration of finer and subtler mechanisms. Some works have also focused on other minority groups (e.g., Hispanics and Asians in US cities) as well as on women. These studies raise interesting research questions on whether and why different groups could be differently affected by spatial mismatch. Are some groups simply less exposed to spatial mismatch or to the effects of spatial mismatch? In other words, do some groups reside closer to jobs or are they simply less affected by distance from job opportunities all things else equal? Are there particular mechanisms that are more relevant for some groups than for others – and why should this be the case? The gender approach to spatial mismatch also raises challenging questions as the location choices of women may be more constrained than those of men and given other gender specificities with regard to more complex commuting patterns, labor market participation, or time schedules (which may also depend on the life cycle of individuals, with activities such as picking up children from school being specific to relatively young individuals). There is also an increasing number of attempts to study spatial mismatch in non-US contexts, especially in European cities (which exhibit spatial structures that differ markedly from US cities) and in developing countries where lack of control over rapid urbanization often results in severe urban sprawl.

Finally, the various analyses of spatial mismatch lead to a diversity of policy implications. Depending on the context and mechanisms potentially at play, policy makers may consider options as diverse as the adoption and implementation of antidiscriminatory laws, the facilitation of residential mobility, neighborhood regeneration policies (in particular through the setup of enterprise zones designed to attract jobs), the development or subsidization of public and private transport, or the spatial dissemination of information on jobs.

In what follows, we present the main theory, empirics, and policy issues surrounding spatial mismatch.

6.2 The Theory of Spatial Mismatch

“Understanding” spatial mismatch requires a focus not only on the labor market mechanisms leading to unemployment, low wages, and poverty but also on what causes ethnic minorities to be physically disconnected from jobs in the first place. Several complementary explanations that can be replaced in a historical perspective have been put forward. They revolve around the (re)location of firms to the suburbs and the reasons why blacks did not move closer to suburban jobs.

The structure of US cities has evolved over the second half of the twentieth century with the emergence of faster and cheaper means of transport for people and goods. A large fraction of middle- and upper-class white workers were able to move to the suburbs to consume more land and build larger houses, as they could commute to inner-city jobs by tramway, train, bus, and, for many, by car. Lower transport costs (resulting from innovations in transportation) also allowed manufacturing firms to relocate to the suburbs to avoid high land prices in the central business district. While many white workers relocated closer to their jobs to

incur shorter commutes while able to increase their housing consumption, the vast majority of blacks did not follow.

When US cities began to decentralize, relocating to the suburbs was an option that was mainly attractive for manufacturing firms as they usually needed a fair amount of land to operate and land was cheaper in the suburbs. The usual agglomeration forces highlighted by economic geography were also at play. As suburban manufacturing activity grew, it fostered the location of firms producing intermediary inputs so as to facilitate the input–output linkage. Services firms providing services to other firms as well as to workers (e.g., convenience services for local households) also followed. More generally, the creation and relocation of firms was also facilitated by the existence of the labor pool consisting of workers located in newly created residential areas. Some firms were attracted to the suburbs by the prospect of benefiting from newly adopted innovations, while others were driven away from city centers because the intensive use of private vehicles had caused congestion problems and because the relocation of firms employing low-skilled labor out of popular neighborhoods had increased the level of unemployment, poverty, and consequently criminality in inner cities. Even firms which had chosen to remain centralized later tended to relocate when criminality reached a tipping point. This “flight from blight” further reinforced the cumulative process of suburbanization.

The explanation initially provided by Kain for blacks not relocating to the suburbs was that blacks faced racial discrimination in the housing market, causing the residential separation of blacks from whites and, indirectly, from suburban jobs. Housing discrimination was indeed certainly a powerful force that shaped US cities in the 1960s when the intuition of spatial mismatch theory emerged and has remained an important driver of segregation. The prevalence of housing discrimination in US cities was unambiguously demonstrated through controlled experiments that assessed the lower number of houses shown by real estate agents to black clients in comparison to the number of houses shown to white clients with similar socioeconomic background (see Yinger 1986). Other studies stressed that discriminatory practices may in fact occur at different stages of the residential mobility process, including during house hunting, borrowing (for those acquiring a home), and rental lease agreement or contract settlement. Mortgage and credit institutions in particular could be applying stricter lending criteria to minorities, constraining their location choices and making their suburbanization more difficult (Ross and Yinger 2002). Interestingly, there can be various motivations underpinning these discriminatory practices, ranging from sheer prejudice (which includes so-called customer discrimination by real estate agents who believe that selling houses to blacks will make the neighborhood less attractive to future white customers) to statistical discrimination from lenders (whereby minority members are expected to have a higher default rate on average).

It is important to understand that although housing market discrimination was initially presented as a key element of the spatial mismatch hypothesis, it is not needed at all to account for the physical disconnection of minorities from jobs. In fact, spatial mismatch can also occur under free location choices according to a variety of mechanisms. In standard land use models in urban economics,

households compete for land, and spatial sorting according to income is a spontaneous equilibrium outcome: As heterogeneous income groups make different trade-offs between proximity to job centers and housing consumption (land being endogenously cheaper further away from places of employment), this may cause the poor – and for historical reasons the minority groups – to live further away from jobs. Separation from jobs may also occur because of the spatial sorting of households in homogenous jurisdictions: As whites and blacks may have different preferences for public goods, they could end up segregating themselves from one another by voting with their feet. This can result in blacks living in inner cities while whites reside in the suburbs where many entry-level jobs are located. Finally, some authors have also put forward (and empirically assessed) the preferences of ethnic groups to live together. This encompasses both mechanisms of white flight (whites seceding from mixed neighborhoods) as well as ethnic clustering of minorities who may want to live together (Ihlanfeldt and Scafidi 2002), even at a distance from jobs.

A number of policies and regulations may also have voluntarily or involuntarily contributed to the disconnection of minorities from jobs. This includes the implementation of most housing projects in city centers (where minorities already live) and in places where land prices are cheaper and that are thus likely to be distant from jobs (Kain 1992). Local zoning regulations in the suburbs may also impose stringent minimum requirements for dwellings as in particular minimum lot sizes, with the implicit objective to prevent an inflow of poorer households to these areas by making housing too expensive for them (Squires 1996). For fear of crime, residents in suburban areas often oppose public transport extensions linking poor areas to their neighborhoods of residence. This further contributes to isolating inner-city minorities from suburban jobs.

There are at least five theoretical mechanisms that can make the distance to job opportunities harmful, especially for ethnic minorities (see Gobillon et al. 2007 for a full description of the corresponding models).

Mechanism 1. The first mechanism relies on *commuting costs* associated with job offers. When a worker receives an offer for a job located far from his place of residence, he anticipates that he will have to incur daily commuting costs if he accepts the offer. These costs can be important enough to outweigh the benefits from even a well-paid suburban job, in which case the worker will turn down the offer. He may prefer to remain unemployed or occupy a lower-wage job which is located closer to his place of residence. This mechanism is particularly relevant for ethnic groups which are not wealthy enough to purchase a car and to pay for its insurance and maintenance and who thus have no other choice than to rely on inefficient public transport.

Mechanism 2. Distance to jobs can also be harmful to workers because it decreases their job *search efficiency*. When searching for a job, a worker may have very little information on which places have suitable job offers and may end up looking for a job in the wrong locations. For low-skill services jobs in particular, the recruiting methods of employers are often local (e.g., ads in local newspapers and “wanted” signs), which may further reduce the information that applicants have on distant job offers.

Mechanism 3. Another mechanism revolves around the idea that *job search costs* can be large and may deter workers from looking for a job in places that are distant from their residence. Job seekers may restrict their search to their neighborhood or its vicinity even if job opportunities in those places are scarce. This is particularly true for workers who do not have a car and depend on inefficient public transports to search for a job in distant places.

Mechanism 4. Workers who reside in areas that are far from job centers and where housing is more affordable may have less incentive to actively look for a job. As a consequence, they may not exert much *job search effort*. Since their housing expenses are lower, they can afford to remain unemployed for a longer period of time than households living in less affordable areas that are closer to jobs. On the contrary, unemployed workers leaving in areas where rents are expensive may feel more pressured to intensively search for a job in order to avoid having to move out.

Mechanism 5. Finally, employers may consider that long commutes deteriorate the *productivity of workers* and may decide not to hire workers who reside too far from the workplace. The reason why productivity may be deteriorated by distance is that distant workers are more likely to be late or tired. This is particularly true for workers located in poor suburbs that do not have a car and use unreliable mass transit.

Several comments can be made about these mechanisms. Interestingly, the consequences of spatial mismatch do not percolate through mechanisms that directly involve ethnicity but rather through the residential location of ethnic minorities within metropolitan areas. In fact, these general mechanisms may in theory apply to any worker who is distant from job opportunities, irrespective of ethnicity. Of course, it does not mean that race does not play a role at all as discussed in the previous subsection on the causes of spatial mismatch. In fact, race can and does play a key role in several respects. First, spatial mismatch can add up to other mechanisms that prevent the employment of minorities in the suburbs such as customer discrimination in fast food restaurants (see Ihlanfeldt and Young 1996) and more generally in suburban services jobs that require contact between clients and employees. The idea here is that, when filling those jobs, employers discriminate against minorities to satisfy the racial preferences of their customers. In this context, residential segregation leads to labor market discrimination (although it should be noted that this involves more the *disconnection* between the neighborhoods than the *distance* between the neighborhoods). This in turn shuts off the access of many suburban jobs to black applicants. Second, spatial mismatch may be all the more relevant in situations of ethnic discrimination in the labor market. The idea is that when minorities are discriminated against, they become more dependent on physical proximity to job opportunities to find jobs (Selod and Zenou 2006). It is also noticeable that these spatial mismatch mechanisms can play at different stages of the job match process and involve both the workers' perspective (for the first four mechanisms) and the firms' perspective (for the fifth mechanism). Finally, even though the spatial mismatch theory focuses on the effect of distance and not on the effect of other neighborhood or group characteristics on labor market outcomes, the

above five mechanisms can also indirectly be amplified by local or group interactions. For example, distance to jobs can have a direct negative effect on workers employment through either one of the five above mechanism, but also indirectly through a feedback process involving localized social network. When most individuals in a location are harmed by distance and are more likely to be unemployed, the local social network is of bad quality, implying that neighbors cannot be used as referrals to potential employers.

6.3 The Empirical Tests of Spatial Mismatch

During the first decades during which the spatial mismatch literature unfolded, most empirical studies aimed to provide some *general test* of the spatial mismatch theory for US cities by assessing whether differences in labor market outcomes between blacks and whites could be related to differences in physical disconnection from jobs. Although establishing causality was usually not done properly, more convincing empirical tests have been proposed over time. Three main strategies have emerged:

- (a) The first strategy is to *instrument the disconnection from jobs with specific local variables* related to local development or industrial composition. This makes it possible to consider only exogenous sources of variations in disconnection from jobs to measure its effect on labor market outcomes. Adopting this strategy, Weinberg (2000) for instance studies the effect of the relative centralization of blacks compared to whites on the black-white employment differential for young workers in large US metropolitan areas. The centralization of blacks is instrumented with historical features of the housing stock and past black centralization. It is found after instrumentation that the larger centralization of blacks relative to whites accounts for around half of the black-white employment differential. Alternatively, Weinberg (2004) focuses on the effect of job decentralization on the black-white employment differential and instruments job decentralization with the city industry composition. Job decentralization is shown to have a negative effect on the employment of blacks relative to whites.
- (b) Secondly, *natural or controlled experiments* can also help address the reverse causality issue (i.e., the fact that it could be the adverse labor market outcomes that cause minority workers to live far from jobs). The idea is to find a subpopulation of workers whose place of residence was determined irrespectively of proximity to job locations. Several papers restrict their analysis to young adults residing with their parents as they have not chosen their location. However, this approach is imperfect as the unobserved characteristics of these young adults may be correlated with those of their parents and therefore be related to residential location. Alternatively, some housing policy measures may in fact render the location of a targeted subpopulation exogenous. To our knowledge, no such experiment has been studied in the USA but some European countries provide a relatively adequate background. In France, for instance, one may choose to restrict the analysis to workers in the housing

public sector considering that applicants cannot choose the precise location of their dwelling which is attributed by public authorities. However, this remains an imperfect strategy given that, in practice, applicants are given the option to decline housing offers and wait for more suitable ones, at least in the beginning of the process. This obviously makes room for some degree of residential choice as a function of local job availability. In Sweden, the spatial allocation process of political refugees in the 1990s provides an interesting and robust framework to study spatial mismatch (Aslund et al. 2010). In the Swedish context, political refugees were indeed dispatched throughout the territory based only on their observed characteristics in applications. This was done without any interaction with public officers, thus making it possible to evaluate the causal effect of job density in the refugees' areas of residence on employment. In the paper, the econometric specifications take into account the observable characteristics that are reported in applications so as to neutralize the possible effect of sorting across space. The results support the role of the disconnection from jobs on employment.

- (c) A third and last approach consists in conducting a *sensitivity analysis* by simulating the extent to which the location choice may be endogenous (Harding 2003; Dujardin et al. 2008). This makes it possible to deal with the endogeneity bias in studies that try to relate the unemployment status of an individual to a neighborhood dummy (which can capture distance from jobs). There is an endogeneity bias if some unobserved individual characteristics affect both the unemployment status and the location dummy. One way to overcome this issue is to simultaneously model the location choice and find an exclusion restriction to identify the effect of location on unemployment. This exclusion restriction consists in having one individual variable explaining the location but having no direct effect on unemployment. However, such an exclusion restriction is hard to find. This problem can be overcome in a sensitivity analysis where one can arbitrarily fix the correlation between the residuals of the unemployment and location equations to a given level and reestimate the model. The results are considered to be robust if the estimated effect of residence on unemployment remains significant for all plausible values of the correlation between residuals.

A few other empirical works have tried to *test some of the five specific mechanisms* whereby distance to job opportunities can affect the labor market outcomes of minorities.

The most famous empirical study is a test of Mechanism 1 above which addresses the role of changing commuting costs following the relocation of a Detroit firm from the city center to a white suburb (Zax and Kain 1996). Whereas whites tended to move closer to the new firm location, it was less often the case of black employees, possibly because of housing discrimination. Following the relocation, the increase in African Americans' commuting distance also induced many of them to quit their job.

Other papers tried to assess the importance of search costs and of lack of information on job opportunities on the bad labor market outcomes of blacks (Mechanisms 2 and 3), although it is usually not possible to distinguish clearly

between the two explanations. For instance, Holzer and Reaser (2000) investigate the application of blacks and whites to jobs in the suburbs using a survey on several metropolitan areas. They find that less-educated black workers apply less frequently for jobs in the suburbs than in the central cities. Evidence provided by Stoll (1999) further shows that increasing blacks' access to cars or decreasing their average distance to search areas would lead them to conduct a more extensive geographical job search.

To our knowledge, Mechanisms 4 and 5 remain largely unexplored by the empirical literature.

Several authors have started investigating whether spatial mismatch could also be of concern for other subgroups of the US urban population. Raphael and Stoll (2001) for instance show that spatial mismatch also contributes to the unemployment of Hispanics and Asians but to a lesser extent than for African Americans. Differences in the vulnerability of the different groups to spatial mismatch point to possible variations across ethnic groups in the level of housing discrimination, in residential location, in access to private and public transports, and in skills (with low-skilled workers more likely to be affected by spatial mismatch). The literature however has not explored the underlying mechanisms so that more detailed studies will be necessary to validate these potential explanations.

While most studies focus on males, recent developments in the spatial mismatch literature have also begun investigating the gender relevance of the theory. Emphasis is put on the residential and workplace location choices of women in multi-person households (which can be tied to those of males or constrained by the presence of children) and on the complexity of the commuting patterns of women which may involve trips to various places such as schools and shops (Blumenberg 2004). This adds complexity to Mechanism 1 above as it is the whole itinerary that is now taken into account in the search and acceptance of job offers, especially for single mothers without cars. To carry out further the analysis, future research could be devoted to studying female spatial mismatch taking into account intra-family decisions (possibly at different stages of the life cycle) and complex itineraries using detailed data on transport patterns.

An important opening of the spatial mismatch literature over the last decade has been to test its validity in several cities outside the USA. Although the historical context and spatial settings in these cities are very different from the USA, this does not preclude a test of the mechanisms. In Europe for instance, urban spatial structure is somehow inverted: Many low-skill jobs tend to be located in relatively central parts of the cities, whereas minorities are residentially concentrated in some relatively peripheral areas. Evidence for these cities, however, is mixed. Among the supportive papers, we already mentioned Aslund et al. (2010) which shows that, in Sweden, the job densities in the places where political refugees are exogenously assigned play a significant positive role on their employment. In greater London, Fieldhouse (1999) finds that employment is correlated with job density for a few ethnic groups, namely, the Pakistani and the Bangladeshi. For Madrid and Barcelona, Matas et al. (2010) show that low job accessibility in public transport negatively affects employment probability. For Paris and Brussels, papers show

that the spatial mismatch hypothesis is not really an issue. Gobillon et al. (2011) find that job density within 45 min by public or private transport is not correlated with finding a job for unemployed workers. Surprisingly, Dujardin et al. (2008) even find a positive correlation between job density and the probability of unemployment. In fact, the evidence in these papers points to vulnerable groups not being largely disconnected from jobs and to segregation effects (in terms of nationality or skill) that are believed to be more problematic than spatial mismatch.

Besides the USA and Europe, there are also many indications of vulnerable groups being physically disconnected from jobs in many other regions of the world, in Latin America, Africa, and Asia (see for instance Sang et al. 2011 on China). In South Africa, this is evidenced by very long and costly commutes. In Johannesburg for instance, the average commute is around 80 min one way, and a national household survey shows that commuters in the poorest income bracket spend about 35 % of their earnings on commuting. Unfortunately and probably due to lack of adequate data, very few studies exist on the effect of physical disconnection to jobs on labor market outcomes in developing countries. South African cities stand as an exception where research suggests a negative impact of distance to jobs on the employment of township residents (Rospabé and Selod 2006).

6.4 Local Policies to Reduce Poverty

After five decades of investigations, the abundant literature on spatial mismatch has shed convincing light on both the causes and consequences of spatial mismatch. Evidence of market failures in both housing and labor markets provides a justification for policy intervention. To categorize the diversity of policy responses in the US case, Ilhanfeldt and Sjoquist (1998) have come up with a useful typology: moving people closer to jobs (desegregation strategy), moving jobs closer to workers (inner-city development strategy), and making it easier for workers to get to existing jobs (strategy of promoting mobility and disseminating information on jobs).

Moving people to jobs is a straightforward recommendation in contexts of constrained mobility. A simple way to address housing, mortgage, and credit markets discrimination is to enforce antidiscrimination policies through the legal system.

Existing public policies could also be modified to facilitate the access of minorities to suburban neighborhoods. In particular, public dwellings could be constructed in predominantly white suburbs with greater job densities. But the policy could prove inefficient in the long run as whites and jobs may respond to the influx of minorities by deciding to move out of these suburbs. The policy would then result in the creation of new deprived neighborhoods out of city centers that may not necessarily be better connected to jobs. Other policy measures could consist in suppressing or forbidding zoning regulations that impose minimum lot sizes in suburbs. This would not necessarily be sufficient though if the constraint is not binding, with developers still targeting rich populations in priority by constructing only large high-quality dwellings.

Another option is to subsidize residential mobility through the granting of rental vouchers (as, e.g., in the Gautreaux program in Chicago 1976–1990 and the Moving to Opportunity program in Baltimore, Chicago, Boston, Los Angeles, and New York 1994–1999). Originally, these experimental programs were meant to facilitate the moving of households out of poor and segregated areas. In particular, a condition to benefit from vouchers in the Moving to Opportunity program was to relocate to a low-poverty neighborhood. However, there has been no assessment of whether these programs helped households get closer to jobs. Whether or not this happened, an assessment of the Moving to Opportunity program shows that it did not lead to a significant improvement of labor market outcomes (Katz et al. 2001). If a similar program were meant to reduce the physical disconnection to jobs, it would need to grant vouchers to households under the condition that they relocate closer to jobs, and the efficiency of such a program would still have to be evaluated. Even if an experimental program had desirable effects at a small scale, it would be difficult to scale it up and a scaled-up policy would probably not be as successful given a general equilibrium effect whereby vouchers could simply end up being capitalized in the housing prices of neighborhoods located close to jobs.

Moving jobs to people has been pursued in a multiplicity of contexts. Inspired from export processing zones, enterprise zones are meant to attract firms in distressed and low job density areas through the provision of fiscal incentives. A key issue is whether jobs in the attracted firms substitute for other local jobs or if the policy is not just displacing jobs between neighboring areas. The effect of the policy may also be limited if local unemployed workers do not have the required skills or if the targeted residential areas simply do not have sufficient space for office development. It transpires from the literature that the evidence on the efficiency of such policies is rather mixed. One drawback of the related studies is that they often focus on the number of firms and gross employment creation rather than on the local level of employment and local poverty in the targeted areas. Nevertheless, there is some evidence in the US case of a significant decrease in unemployment and poverty related to the introduction of the federal Empowerment Zone program. For France, the introduction of the French enterprise zone program has been shown to have only a small positive effect on finding a job for unemployed workers located in the Paris region. Maybe the creation of jobs adapted to workers' skills could be encouraged by providing tax incentives only to firms in some specific activity branches or for specific jobs. This kind of targeting is usually not implemented in the existing enterprise zone programs. Other place-based policies designed to attract firms include measures to decrease criminality as it can affect firm productivity through vandalism and violence of which employees may be victims. Other policies also include investments in transport infrastructures such as connections to highways that can decrease the transport cost of goods.

Improving connections between people and jobs may seem easier than the above-mentioned options. Improving transport will decrease commuting and search costs and increase the search efficiency and productivity of workers (Mechanisms 1, 2, 3, and 5). This can be achieved through improvements in public transport (adding train and subway stations, increasing transport frequency, or subsidizing fares).

In the USA and other places, this would effectively target minorities who tend to have a lower access to cars and use public transit. However, the extension of the public rail network from the city center to the suburbs may be hindered by the opposition of suburbanites by fear of displacement of criminality. Moreover, public transport improvements may also have some drawbacks: The creation of a new station is likely to cause a local increase in housing prices which could in turn induce renters to move away to further locations. It is also hard to improve connections throughout the city for trains and tramways as it could require massive investments. Adding bus stations and increasing bus frequencies should be less costly, but buses are affected by traffic jams. A better access to private transport could also be achieved with the provision of vouchers to purchase motor vehicles. However, increasing the access to private transport should increase the overall traffic and congestion. Moreover, improved accessibility may paradoxically provide little incentive for poor households to move to better locations and may consequently reinforce segregation.

Beyond transport policies, facilitating the circulation of information flows between firm and workers can also help overcome the information hurdles associated with physical distance. Disseminating the information on the spatial distribution of job openings can greatly help job seekers apply to right places. Improving the flow of information can be achieved by creating local employment agencies in poor neighborhoods where they are missing and by better targeting the informational needs of unemployed workers regarding job offers. In particular, local employment centers could organize meetings with suburban businesses to give an opportunity to unemployed workers to meet with potential employers face to face (Ihlantfeldt and Sjoquist 1998).

6.5 Conclusions

Since the end of the 1960s, a large literature has focused on the contribution of spatial mismatch to the bad labor market outcomes of ethnic minorities. The importance of this contribution can be assessed in contrast to alternative explanations. With the benefit of hindsight, we derive three principal lessons from our review.

First, spatial mismatch refers to mechanisms that can apply in many different contexts and it should not be considered as a topic that may be valid only to explain the poverty of African Americans in US inner cities. There are many indications (and in some cases scientifically determined evidence) that vulnerable populations in the USA and elsewhere are affected by similar problems of disconnection between places of residence and places of employment.

Second, although spatial mismatch is a spatial theory of local unemployment, it should be clear that other spatial mechanisms may also contribute to poor labor market outcomes in poor areas. As a matter of fact, residential segregation constitutes a competing spatial explanation to the unemployment of ethnic minorities through a variety of mechanisms (e.g., the existence of local peer effects on

employability, deteriorated information networks on jobs, and discriminating employers using neighborhood composition to infer information and employability). In the current state of research, it is not clear however whether it is spatial mismatch or segregation that contributes the most. In some contexts, only one or the other may play a role. In other contexts, they probably combine and amplify one another. What is established however is that spatial factors largely contribute to and are probably among the main factors explaining economic and social outcomes and in particular local poverty.

Third, there are also important nonspatial factors at play (e.g., sheer labor market discrimination or skill bias) that can explain the unemployment of vulnerable groups. A direct implication is that policies addressing such nonspatial factors may also have an effect by locally alleviating unemployment. Another implication is that place-based policies will, of course, not suffice to solve the unemployment problems of ethnic minorities. In this context, an important challenge for policy makers is probably to find the right policy mix that is needed between spatial and nonspatial policies.

Acknowledgments We would like to thank all our respective coauthors on our work on spatial mismatch for the many interesting discussions that helped us better understand the topic. Readers may find additional insights on the topic by reading the chapters on ► Chap. 3, “Labor Market Theory and Models,” ► Chap. 4, “Job Search Theory,” and ► Chap. 5, “Commuting, Housing, and Labor Markets,” in the present edition of the Handbook of Regional Science. The findings, interpretations, and conclusions expressed in this chapter are ours and do not represent the view of our employers, including the World Bank, its executive directors, or the countries they represent.

References

- Aslund O, Osth J, Zenou Y (2010) How crucial is distance to jobs for ethnic minorities? Old question – improved answer. *J Econ Geogr* 10(3):389–422
- Blumberg E (2004) En-gendering effective planning: transformation policy of low-income women. *J Am Plann Assoc* 70(3):269–281
- Dujardin C, Selod H, Thomas I (2008) Residential segregation and unemployment: the case of Brussels. *Urban Stud* 45(1):89–113
- Ellwood J (1986) The spatial mismatch hypothesis: are there teenage jobs missing in ghetto? In: Freeman R, Holzer H (eds) *The black youth unemployment crisis*. University Chicago Press, Chicago, pp 147–185
- Fieldhouse E (1999) Ethnic minority unemployment and spatial mismatch: the case of London. *Urban Stud* 36(9):1569–1596
- Gobillon L, Selod H, Zenou Y (2007) The mechanisms of spatial mismatch. *Urban Stud* 44(12):2401–2427
- Gobillon L, Magnac T, Selod H (2011) The effect of location on finding a job in the Paris region. *J Appl Econ* 26(7):1079–1112
- Harding D (2003) Counterfactual models of neighborhood effects: the effect of neighborhood poverty on dropping out and teenage pregnancy. *Am J Sociol* 109(3):676–719
- Holzer H, Reaser J (2000) Black applicants, black employees, and urban labor market policy. *J Urban Econ* 48(3):365–387

- Ihlanfeldt K, Scafidi B (2002) Black self-segregation as a cause of housing segregation: evidence from the multi-city study of urban inequality. *J Urban Econ* 51(2):366–390
- Ihlanfeldt K, Sjoquist D (1998) The spatial mismatch hypothesis: a review of recent studies and their implications for welfare reform. *Hous Policy Debate* 9(4):849–892
- Ihlanfeldt K, Young M (1996) The spatial distribution of black employment between the central city and the suburbs. *Econ Inq* 34(4):693–707
- Kain J (1968) Housing segregation, negro employment, and metropolitan decentralization. *Q J Econ* 82(2):175–197
- Kain J (1992) The spatial mismatch hypothesis: three decades later. *Hous Policy Debate* 3(2):371–460
- Katz L, Kling J, Liebman J (2001) Moving to opportunity in Boston: early results of a randomized mobility experiment. *Q J Econ* 116(2):607–654
- Matas A, Raymond J-L, Roig J-L (2010) Job accessibility and female employment probability: the cases of Barcelona and Madrid. *Urban Stud* 47(4):769–787
- Raphael S, Stoll M (2001) Can boosting minority car-ownership rates narrow inter-racial employment gaps? In: Rothenberg Pack J, Gale W (eds) *Brookings-Wharton papers on urban economic affairs 2001*. Brookings Institution Press, Washington, DC, pp 99–145
- Rospabé S, Selod H (2006) Does city structure cause unemployment? The case of Cape Town. In: Bhorat H, Kanbur R (ed) *Poverty and policy in post-apartheid South Africa*, Chapter 7. HRSC Press, Cape Town, pp 262–287
- Ross S, Yinger J (2002) Color of credit: mortgage discrimination, research methods, and fair lending enforcement. MIT Press, Cambridge
- Sang E, Song J, Xu T (2011) From “spatial bond” to “spatial mismatch”: an assessment of changing jobs-housing relationship in Beijing. *Habitat Int* 35(2):398–409
- Selod H, Zenou Y (2006) City structure, job search, and labor discrimination. Theory and policy implications. *Econ J* 116(514):1057–1087
- Stoll M (1999) Spatial job search, spatial mismatch, and the employment and wages of racial and ethnic groups in Los Angeles. *J Urban Econ* 46(1):129–155
- Squires G (1996) Closing the racial gap? Mortgage lending and segregation in Milwaukee suburbs. Study prepared for the Fair Lending Coalition, Institute for Wisconsin’s Future
- Weinberg B (2000) Black residential centralization and the spatial mismatch hypothesis. *J Urban Econ* 48(1):110–134
- Weinberg B (2004) Testing the spatial mismatch hypothesis using inter-city variations in industrial composition. *Reg Sci Urban Econ* 34(5):505–532
- Yinger J (1986) Measuring racial discrimination with fair housing audits. *Am Econ Rev* 76(5):881–893
- Zax J, Kain J (1996) Moving to the suburbs: do relocating companies leave their black employees behind? *J Labor Econ* 14(3):472–504

Francesca Mameli, Vassilis Tselios, and Andrés Rodríguez-Pose

Contents

7.1	Introduction	110
7.2	The Determinants of Regional Disparities in Unemployment Rates	111
7.3	A Simple Model Based on Supply and Demand	112
7.3.1	Labor Demand	112
7.3.2	Labor Supply	113
7.4	Changes in Labor Demand and Productivity	114
7.5	Changes in Industry Composition	114
7.6	Labor Supply Constraints	115
7.6.1	Human Capital and Skills	115
7.6.2	Demographic Factors	116
7.6.3	Barriers to Labor Mobility	118
7.7	Policy Constraints	119
7.7.1	Social Insurance	119
7.7.2	Place-Based Policies That Limit Mobility	121
7.8	Conclusions	121
	References	123

F. Mameli (✉)

Dipartimento di Scienze Economiche e Aziendali and CRENoS, Università degli Studi di Sassari,
Sassari, Italy

e-mail: mameli@uniss.it

V. Tselios

Geography and Environment, University of Southampton, Southampton, UK
e-mail: v.tselios@soton.ac.uk

A. Rodríguez-Pose

Department of Geography and Environment, London School of Economics, London, UK
e-mail: a.rodriguez-pose@lse.ac.uk

Abstract

A prominent theme in the socioeconomic and regional science literature has been the topic of unemployment. We focus on *regional unemployment* and put forward candidate series of explanations for it using a basic model of labor supply and demand. The persistence of regional unemployment differentials points to inefficiencies in labor markets that in the long run could affect aggregate unemployment rates. Both a lack of labor demand and a constraint of labor supply increase regional unemployment. We finally discuss people- and place-based policies which aim to reduce high unemployment rates.

7.1 Introduction

Unemployment is a social, political, and economic plague that affects modern economies. It generally reduces national economic growth, increases inflation, favors the inequality of income distribution, and also carries important *human consequences*. People decide to be employed not only to earn income for living but also, to a certain extent, for nonpecuniary reasons. Individuals want to work to enjoy the feeling of doing something productive, of being needed, of reaching a certain social status, etc. Though unemployment increases leisure, the value of this may thus be wholly offset by the feeling of rejection, and the new status could have profound effects on both the mental and physical health of individuals. Further, from a labor market perspective, there is a risk that the unemployed lose some of the skills or human capital previously possessed, making it increasingly difficult for them to find employment in the future. Employers also tend to assume that those who have been out of the labor market for a long period are not as qualified or reliable as those who have been working more recently. This may result in a number of dissatisfied individuals that remain unproductive, which could ultimately decide to permanently leave the labor market, thereby affecting an economy's growth potential.

Unemployment rates vary considerably across regions. The persistency of regional unemployment differentials is a symptom of inefficiencies in labor market adjustments that in the long run could affect aggregate unemployment and total output. Understanding the determinants of spatial variation in unemployment is crucial for adopting the appropriate policy instruments able to reduce these disparities and limit the adverse effects of unemployment in economically depressed regions.

This chapter provides a review of the theoretical and empirical literature on regional employment and unemployment. In Sect. 7.2 we discuss the causes of spatial differences in regional unemployment rates. Section 7.3 sketches a simple model of supply and demand of labor. Section 7.4 examines how a lack of labor demand and productivity shortfalls may limit regional employment. Section 7.5 considers the effect of industry composition on unemployment disparities. Section 7.6 analyzes how labor supply is constrained by human capital and skills,

demographic features, and a lack of mobility that limits movements to expanding regions. Section 7.7 discusses the effects of unemployment social insurance and place-based policies. The final section concludes.

7.2 The Determinants of Regional Disparities in Unemployment Rates

The nature and persistence of regional unemployment differentials have attracted growing scholarly attention since the late 1960s (see, for instance, Thirlwall 1966), giving rise to a large number of potential explanations for the existence of such disparities. These may be broadly ascribed to three categories: (i) “*labor supply*” factors, such as the composition of regional labor force in terms of age, gender, and ethnicity, or the average level of education attainment and skills possessed by individuals (these features affect the regional labor force participation behavior and workers decisions to migrate in search of better work opportunities); (ii) “*labor demand*” factors, which point at differences between goods markets as determinants of an uneven labor demand (within this type of explanation, all factors affecting location decisions by firms, such as the different regional specialization patterns or the economic influence played by surrounding, geographically contiguous regions, are also included); and (iii) “*the flexibility of wages*,” which may be reduced, for instance, by the existence of minimum wages or union activity. If labor markets were efficient, the adjusting forces of labor and capital mobility and changes in relative prices would eventually eliminate unemployment differentials between regions. On the other hand, while internal migration may act as an adjustment mechanism to reduce general unemployment dispersion, it also varies considerably across regions due to the existence of mobility barriers which may limit movements to expanding regions. The decision to move is indeed complex and affected by multiple factors such as personal attributes, cultural reasons, individual labor market situations, the high costs involved with moving, or the generosity of unemployment insurance.

From a theoretical perspective, two competing explanations account for the existence of unemployment differentials between regions (Marston 1985). The first one is that there is an equilibrium relationship of unemployment rates across areas. Workers migrate in search of better work opportunities until there is no further incentive to move because they feel somehow compensated (e.g., by local amenities and land endowments). Each region tends to its own *equilibrium unemployment* rate and the existence of persistent unemployment disparities between regions simply reflects the underlying preference of workers for some areas. Contrary to the equilibrium view, the *disequilibrium explanation* assumes that labor flows slowly between areas because of severe economic and social barriers restricting mobility, which generate persistent unemployment rate differentials between regions. The equilibrium interpretation has received empirical support from Marston (1985) – who finds that high unemployment rate areas are those with high wages, high unemployment insurance, and attractive amenities – and Partridge and Rickman (1997).

Over the years, the empirical relevance of the relative importance of these explanations has been scarce. This is probably because the equilibrium and disequilibrium views on long-term regional disparities are not mutually exclusive and a model allowing the conterminous testing of both theories is considered a near impossible task (Pehkonen and Tervo 1998). By contrast, the analysis has been more oriented toward identifying which determinants of the spatial variation in unemployment rates have been more important.

7.3 A Simple Model Based on Supply and Demand

In order to provide a framework for the analysis of the labor market, this paragraph sketches a basic *model of supply and demand* (Johnson and Layard 1986) that defines the equilibrium long-run unemployment rate at which the system would settle down if prices and wages were correctly foreseen. In this model (depicted in Fig. 7.1), at given real wage and labor force, only a fraction of the labor force (L) wants to work. If the market clears, unemployment is simply leisure, voluntarily chosen.

7.3.1 Labor Demand

Firms have the following aggregate production function with constant returns to scale:

$$Y = F(N, K) = Nf(k) \quad f' > 0, f'' < 0 \quad (7.1)$$

where Y is output, N is employment ($= LE$, size of employed labor force), K is capital, and $k = K/N$. The marginal productivity condition is

$$W = F_N(LE, K) \quad (7.2)$$

In the short run, capital is given (\bar{K}) so that labor is the only factor. Firms demand as much labor as it is necessary to equate the marginal productivity of labor to the real wage and the demand curve is downward sloping.

In the long run, if the real interest rate is fixed, the marginal productivity condition under constant returns to scale is:

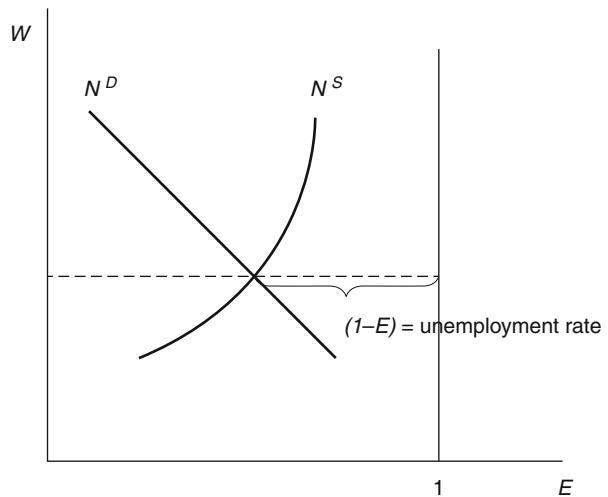
$$W = f(k) - kf'(k) \quad (7.3)$$

and

$$r = f'(k) \quad (7.4)$$

If r is given, the demand price for labor depends on Eqs. (7.3) and (7.4).

Fig. 7.1 Supply and demand model of unemployment, market-clearing



7.3.2 Labor Supply

On the supply side, we assume that individuals differ in their taste for work, with some individuals becoming permanently unemployed and others remaining permanently employed. The supply of labor is determined within the context of an individual's work-leisure choice. Utility (U) depends on consumption (C) and hours of work (H). Individuals can choose between working a given number of hours per week (\bar{H}) and receiving a wage W , or not working at all ($H = 0$) without income ($Y = Y_0$). The i th person utility is given by:

$$U_i = U_i(C_i, H_i) \quad (7.5)$$

She is indifferent between working or not (i.e., unemployment) at a reservation wage W_i^* , so that:

$$U_i(W_i^*, \bar{H}) = U_i(Y_0, 0) \quad (7.6)$$

The individual maximizes utility given his budget constraint. The resulting supply decision will depend on his real wage (W) and other parameters of his budget constraint (Z):

$$E = E(W, Z) \quad (7.7)$$

where E is the employment rate, which is one minus the unemployment rate. The attractiveness of entering the labor market may depend on various factors such as, for instance, the tax system, the generosity, and the duration of unemployment benefits. For a given Z , the supply curve is a positive function of real wages and is

upward sloping (starting from the lowest W^*). In the absence of taxes and benefits, the equilibrium unemployment rate results from the intersection of the demand and supply curves and is equal to one minus the employment rate.

7.4 Changes in Labor Demand and Productivity

From a Keynesian perspective, *labor demand* is mainly constrained by the level of aggregate demand (the demand for labor is in fact considered a “derived demand”) and unemployment is a direct result of demand deficiencies in the product market. A reduction in demand for the region’s goods and services causes an inward shift of the labor demand curve that, in presence of sticky real wages, reduces the number of people actually employed while maintaining the number of those that wish to work. Basically, workers become (involuntarily) unemployed if the demand for firms’ output is not enough to induce employers to take on all the labor actually available for employment.

Shifts in regional labor demand may be caused by persistent *productivity changes*, which can vary across regions depending upon multiple factors, such as their industrial composition, investments in technology, or endowments of human and physical capital. When productivity decreases, labor demand may contract because wages cannot decrease (they are sticky), causing involuntary unemployment. A productivity slowdown may increase unemployment via two mechanisms acting on wages’ inertia: (a) workers’ wage aspirations do not readily adjust downward and (b) unions push for sudden higher wages.

7.5 Changes in Industry Composition

From the discussion above, it could be expected that the different *industrial mix* of regional economies is a crucial factor affecting labor demand and, hence, unemployment. The available empirical evidence is however controversial. There are some studies that consider the industrial composition a major explanation for divergent unemployment rates across areas. McGee (1985), for instance, argues that regions with a larger proportion of the labor force in industries with high (low) unemployment rates are considered more likely to have higher (lower) unemployment rates than the national average. Other research suggests instead that spatial differences in industrial composition account for little, if any, of the variation in unemployment rates across regions (Cheshire 1973; Taylor and Bradley 1983; Martin 1997). Martin (1997), for instance, finds that the same industry experiences different unemployment rates in different regions, with a strong tendency for unemployment to become equalized between industries within each area (because high unemployment industries exert a negative spillover effect on other industries in the area). Taylor and Bradley (1997) suggest that the effect of industry mix on unemployment disparities varies not only between regions but also between sectors. They find that Italian regions with a high proportion of workers in the

manufacturing sector tend to have lower unemployment rates, while having a high proportion of workers in the agricultural sector is particularly disadvantageous. The opposite is true for the UK, where the regions more heavily dependent on agriculture are those that have lower unemployment rates. High dependence on market services in UK regions is in turn correlated to substantially lower unemployment rates. On the whole, the observed heterogeneity in results suggests that large differences in employment performance among regions may not be explained exclusively in terms of disparities in industry composition. They probably depend also from other characteristics, such as the extent of innovative activity, the profitability of plants and firms, the degree of accessibility to product and factor markets, and the effects of agglomeration (dis)economies.

7.6 Labor Supply Constraints

7.6.1 Human Capital and Skills

The supply of labor depends on investments in *human capital* and *skills* attained by workers, which influence labor market attachment of the labor force. The different composition of regional population in terms of these factors, combined with a global skill-biased demand shift, may partly account for spatial variations in unemployment rates across regions.

The level of educational attainment within a region is one potential indicator of the stock of regional human capital and is a proxy for skills and matching success in local labor markets. Higher levels of education tend to be positively related to labor force participation rates and negatively related to unemployment rates (McGee 1985; Partridge and Rickman 1997). Employment rates of less-educated people are also much lower than those of the more educated. Less-educated people have a limited access to the labor market and are unlikely to find work even if there is an increase in labor demand, because they do not possess the skills, or are in some way unsuitable, for the jobs on offer. To give an example, the unemployment rate in the Euro area in 2007 was 9.3 % for those with lower secondary education, 6.4 % for those with upper secondary education, and 4.1 % for those with tertiary education.

In general, *unskilled* people have much higher *unemployment rates* than those with skills, which tend to be more selective and also demonstrate a preference for employed search. The *vulnerability of low skilled* may depend on different factors (Metcalf 1975): (i) they lack information; (ii) skilled workers have more firm-specific training and (in case of recession) employers tend to hoard skilled labor or, at least, decide to lay off the unskilled workers first; and (iii) they cannot compete with skilled workers for many jobs, while skilled workers are also able to perform unskilled jobs. In addition, due to new technologies and international competition, many countries have experienced a dramatic change in the structure of employment in recent years. This has shifted production/demand requirements toward workers with higher qualifications and higher skill levels, leading to a collapse of demand

for the unskilled. Consequently, regions hosting a high proportion of low-skilled workers may have seen their unemployment rates rise (Overman and Puga 2002).

Different migration patterns characterize individuals with different levels of educational attainments and abilities as the better educated, more skilled, and adaptable workers tend to be also the most geographically mobile (see, e.g., Courchene 1970; Pissarides and Wadsworth 1989; Martin 1997; Partridge and Rickman 1997). There are a number of potential explanations for such *self-selecting migration*: (i) skilled professional workers are more able to meet the costs of moving and also have better chances to receive assistance from employers to meet these costs; (ii) individuals with a higher level of human capital are more likely to migrate to obtain higher returns; (iii) low-skilled people may rely more on friends and families for support (e.g., transportation or child care), making it risky for them to move; and (iv) the psychological reluctance to moving may be lower for well-educated workers, especially if they pursued advanced studies away from their homes.

As suggested by Martin (1997), labor migration of the most educated-skilled-adaptable workers may perpetuate regional unemployment differentials by transforming temporary regional labor demand problems (or an unfavorable industrial composition) into long-term, region-specific, labor supply problems. The skills and benefits associated with enterprising workers may be lost to a region, and its skill base and work culture can become cumulatively less viable and less productive, which in turn may militate against employment growth and hence any substantial reduction in unemployment (Martin 1997: 246).

We want to stress that an increase in employment is not directly translated in a decrease of unemployment because the relationship is not one-to-one. Employment growth at the regional level may not reduce the regional unemployment rate because in-migrants may absorb all the new jobs leaving unemployment unaffected (Partridge and Rickman 1997). As suggested by Blanchard and Katz (1992), “trends in employment do not lead to trends in unemployment” (p. 29).

7.6.2 Demographic Factors

The labor supply is constrained by *demographic factors* like age, gender, and ethnicity, which influence the number of people available and willing to participate in paid employment. Changes in demographic features of labor supply may have considerable consequences for the wage structure and unemployment rates of sub-groups of workers. Assuming flexible wages and unchanged labor demand, an increase (decrease) in the relative supply of a specific type of worker results in a lower (higher) wage relative to other types of workers. If, however, employers do not regard workers from various groups as good substitutes for one another and wage rigidities occur (at least in the short run), this may result in a mismatch between unemployed workers and available jobs (i.e., an increase of structural unemployment). On the whole, when some demographic groups register above-average unemployment rates, the aggregate unemployment rate increases (because

the overall rate is a weighted average of the rates of all groups). Regional differences in (un)employment may therefore be driven by regional differences in the composition of the labor supply and participation trends.

7.6.2.1 Age and Gender Differences

If we look at the characteristics of the different demographic groups, on average *young people* are more likely to be unemployed than older people; they enter/exit the labor force more often than adult men and exhibit weaker job and labor force attachment (McGee 1985). This is possibly explained by the fact that they lack skills and experience with respect to older workers. In most countries, they also work in different sectors from adults with a tendency to find more occupations in the retail, hotel, and restaurant industries rather than in education, utilities, or the public administration. Their successful transition into the working world varies considerably by education attainment in every country, with higher unemployment rates for those with the least education. Similarly, the labor market participation of the elderly tends to be very low. On the whole, however, the empirical literature suggests that regions with an older population register lower unemployment rates, while the opposite occurs when there is a large proportion of young labor force.

Turning to *gender differences*, women have often a limited access to the labor market, though there are cross-national differences. In the European Union, for instance, participation and *employment rates* are substantially higher for males than for females in all age groups. The causes of gender inequality in the European labor market are quite complex, with a variety of political, administrative, and legislative responses implicated. Women have more responsibilities for caregiving and household tasks than their male partners. Many women, particularly those who are heads of households with young children, have a tendency either unemployed or limited in their employment opportunities for reasons that include inflexible working conditions and arrangements, inadequate sharing of family responsibility, and a lack of sufficient services such as child care. Many women stop working altogether after their having their first child, while others return to the labor market as part-time workers or when their child or children are of school age (Rodríguez-Pose 2002: 80). The cultural barriers, including the persistence of informal networks from which women are excluded, also prevent them from achieving equal participation in the labor market. The effect of women's individual characteristics which shape their access to the labor market may depend on the sociopolitical structure, such as the male-dominated hierarchy of the political economy and the existing ideologies on gender. According to Barnes et al. (2005: 171), gender inequalities at the regional level reveal the predominance of women in part-time work, women's underrepresentation in sectors such as engineering. Rodríguez-Pose (2002) argues that there is an age and gender divide in atypical employment forms, because the number of women working part-time is higher than that of men, whose part-time employment is concentrated among the young and the over 55s, while self-employment is basically a male phenomenon. People with lower skills are being relegated to these forms of employment and condemned to lower salaries. The concern is whether the differences in access to work for different age groups and the

gender divide in employment can be justified by intergroup differences in worker attributes or whether these differences are the result of employment discrimination and unfair access to work.

7.6.3 Barriers to Labor Mobility

The elasticity of labor supply largely depends upon the geographical mobility of labor: the more readily labor can transfer between regions, the more elastic the supply curve. When mobility increases, geographical mismatches in labor supply and demand are more easily accommodated, and consequently unemployment may be low. Conversely, a more immobile work force may result in higher structural unemployment due to an insufficient supply of labor.

A wide array of individual considerations may affect mobility decisions of workers by making them more attached to their place of residence. These include, for instance, the anxiety of relocation; the unwillingness of people to leave their social, cultural, and family ties (i.e., psychic costs); a lack of information about jobs in other areas; or the pecuniary costs involved with moving (e.g., costs of selling a house). On the whole, mobility propensity is largely influenced by personal characteristics of movers (e.g., age, education, expectations), but in general workers decide to migrate if the expected utility of moving is higher than the expected utility of staying, net of migration costs. Human capital theory in particular sees mobility as an investment decision in which monetary and psychic costs are undertaken in an initial period in order to obtain returns over a longer period of time. The choice of migrating is based upon the net present value of the costs and benefits of such an investment (Becker 1962).

Housing market prices and tenure choices are frequently cited as an obstacle to geographical labor mobility. The empirical evidence shows that high relative price houses may prevent labor market adjustment by discouraging net migration to a region. Potential movers may in fact find that the housing price differential is too large an obstacle to overcome. Moreover, mobility patterns are strongly affected by housing tenure choices. Those living in rented public-sector housing (i.e., council houses) are less likely to move regions than other types of tenants, while owner-occupiers are less mobile than private-sector renters (Hughes and McCormick 1987).

There are a number of possible reasons for the relative immobility of council tenants. One is that public renters wishing to migrate are expected to rely upon council house exchange/transfers, but it is possible that the number of households seeking to enter and leave an area may be unequal (Hughes and McCormick 1987). Another explanation lies in the allocation procedure of properties, which often gives high weight to residence in the local authority, thereby increasing the difficulty of moving into equivalent subsidized houses in a different area.

On the other hand, the reduced mobility of home owners relative to private renters largely depends from the high search and transaction costs involved with buying and selling a house (e.g., transaction taxes, such as capital gain taxes and ad

valorem taxes which are proportional to the house value) which makes them more location bounded. Consequently, as transaction costs are spread over a longer period, most households decide to buy a house the longer the expected length of stay in a dwelling. For the same reason, mortgage holders are more likely to make a long-term locational commitment and are the least mobile group, especially when the mortgage interest rates are high.

The mobility constraints imposed by council housing and home ownership hinder the propensity of individuals living in these types of tenancies to move from regions experiencing a decline in labor demand to expanding regions. Based on this finding, Oswald (1997) suggests that home ownership represents a major barrier to smoothly operating labor markets, which increases *unemployment disparities* between regions. Numerous authors have tested this prediction with different regional data and most aggregate analysis confirms a positive relationship between owner occupation and unemployment (e.g., Partridge and Rickman 1997). At the microlevel (controlling for different characteristics of individuals or households), the evidence is however less favorable to *Oswald's hypothesis*. Homeowners have not only lower unemployment probabilities but also higher wages and shorter spells of unemployment than renters (Coulson and Fisher 2002). With respect to council housing, the literature confirms a positive correlation between the regional unemployment rate (or the probability to be unemployed) and the proportion of households in the public housing sector (Taylor and Bradley 1983; McCormick 1983).

7.7 Policy Constraints

7.7.1 Social Insurance

Public *unemployment insurance* (UI) is a social institution meant to provide temporary financial assistance to workers who are displaced from their job. Despite operating beneficial redistributive effects on those suffering an income loss, thereby reducing intrapersonal income disparities, the unemployment insurance adversely affects labor market behavior. In those cases where unemployment benefits are generous, they may contribute to raise the unemployment equilibrium and to reduce the speed of labor market adjustment after an exogenous shock. They may also reduce the rate at which displaced workers become employed in other sectors. In particular, four features of the unemployment benefit system influence equilibrium unemployment: the level of benefits, the maximum duration of entitlement, the coverage of the system, and the strictness with which the system is operated.

Since the early 1970s, a large number of studies have investigated the effect of unemployment insurance on unemployment level and duration. The main predictions for the *effects of UI* are mainly derived from job search theory, which provide rigorous and detailed analysis of rational individual behavior during unemployment. The standard result from search theory is that unemployment benefits are expected to raise beneficiaries' reservation wages, thereby reducing their search

efforts and their willingness to accept job offers. Workers have perceptions of the wage distribution and voluntarily decide to remain unemployed when current wages are lower than their reservation wage, increasing optimal search time. Reducing the opportunity cost of rejecting a job and prolonging the unemployment spells, UI (among other factors) may inhibit geographical mobility by favoring a stronger geographical attachment. This argument is consistent with the typical negative correlation between mobility rates and the generosity of UI in Europe and the United States: Europe is characterized by high UI and low mobility, while the USA features low UI and high mobility.

Research seems to confirm that the more generous unemployment benefits, the longer the unemployment spells and the higher overall unemployment rate (see, e.g., Moffitt and Nicholson 1982; Marston 1985). Over the period 1984–1997, for instance, Partridge (2001) shows that a one standard deviation rise in average provincial UI in Canada increased the unemployment rate by about 1.3 points.

Regional differences in unemployment insurance payments are in particular suspected to “short-circuit” the market forces that would naturally induce migration from low- to high-income areas (Shaw 1986). Assuming that migration decisions depend exclusively on earned incomes or labor productivities in any location, regional differences in the amount of *unemployment insurance* benefits would have no direct *effect on migration*. Workers would tend to move toward regions where they could earn a higher income. However, as they compensate covered workers for a large fraction of the income lost for being unemployed, UI benefits affect the expected income of individuals in the areas where they live (Fields 1979). The result is that higher benefits may subsidize residence in regionally depressed regions by adversely influencing potential migrants’ decisions, therefore prolonging unemployment and contributing to maintain inequality in unemployment rates among regions. Nonetheless, the effect of UI depends in general on the design features of the unemployment benefit system (in terms of eligibility conditions, level and duration of benefits) and the availability of effective active labor market policies that stimulate job search but do not create obstacles to mobility (OECD 2005).

Empirical research has documented that UI has had an adverse effect on Canadian interregional migration patterns. Courchene (1970), for instance, demonstrates that the generosity of unemployment insurance payments in Canada, relative to the USA, had a significant negative effect on out-migration in Canadian provinces over the period 1952–1967. Similarly, Winer and Gauthier (1982) report that out-migration from the high unemployment Atlantic provinces has been inhibited by increases in UI benefits in the Atlantic region and stimulated by increases in UI support in other regions. The evidence for other countries is weaker. For the United States, Fields (1979) finds that the amount and availability of UI benefits have had some effect on the geographic mobility of labor, but these effects are not very large. Goss and Paul (1990) show that UI compensation decreases the migration likelihood only for those who are involuntarily unemployed, while the opposite occurs for the voluntarily unemployed. For Europe, Tatsiramos (2009) observes that receiving benefits is not associated with lower mobility, but the effects might

vary depending on the institutional characteristics and incentive structure of the UI system of each country. In countries with relatively generous benefits (Denmark, France, and, to some extent, Spain), recipients of UI are more likely to move compared to non-recipients, while in countries that provide less generous benefits (UK), or generous but with a relatively weak incentive structure (Germany), recipients do not differ significantly on their likelihood to move.

7.7.2 Place-Based Policies That Limit Mobility

Differences in (un)employment rates across regions may also arise in reaction of “place-based” policies aimed at helping economically disadvantaged areas/communities, which may alter the adjustment of regional labor markets. These policies take the form of subsidies, grants, or tax incentives to local economic development of specific areas.

An example of *regional spatial policies* with a particular emphasis on worklessness is the *enterprise zones*. These programs were initially developed in the UK in the early 1980s to induce property development as well as industrial and commercial investment in selected areas by removing/reducing certain fiscal burden (mainly local taxes and taxes on capital investment), simplifying administrative procedures and reforming certain statutory controls like planning regulations (OECD 2005). Similar policies have been adopted in other European countries and the USA, which also launched an Empowerment Zone program with similar purposes. Although intuitively appealing, evaluation of enterprise and empowerment zones is rather controversial in terms of social outcomes and it is questioned whether the attained benefits in terms of job growth are sufficient to balance their costs (Glaeser and Gottlieb 2008).

Despite having always been politically popular, economists have traditionally expressed little support for place-based policies in the employment realm, fearing they would alter economic behavior by sending incorrect signals about where to live and to work. Opponents to place-based policies claim that these create *spatial distortions* that limit the tendency of the poor to exit declining areas for localities that offer better employment opportunities (Glaeser and Gottlieb 2008). This is especially the case for low-skilled people, which risk remaining trapped in places with a limited future. In general, standard models of spatial equilibrium are against place-based policies, suggesting that subsidies to poor places are anyhow offset by higher prices and arguing that their primary real effect can be just to redistribute jobs to economically unproductive areas (Glaeser and Gottlieb 2008).

7.8 Conclusions

Researchers across a wide range of fields, policy makers, and large segments of the public believe that unemployment has important social, political, and economic consequences, especially for those regional economies subject to persistently high

levels of joblessness. High regional unemployment means that regional economies not only are leaving talent and skills untapped – thus reducing regional economic growth and development – but may also lead to social unrest, as well as to greater misery and hopelessness for those affected (Begg et al. 2008). It is widely known that putting demand and supply together, we can determine equilibrium prices and quantities in different factor markets, i.e., labor, capital, and land. Each firm simultaneously decides how much output to supply and how many factors to demand. The demand and supply curves are inextricably interlinked. But are these markets always in equilibrium? Are there regional disparities in factor markets and in their returns? In this chapter, we have discussed the regional labor supply and demand and the determinants of regional disparities in unemployment rates. Each regional market tends to clear at the equilibrium wage and employment level that equates supply and demand. However, it may not be possible to take labor market equilibrium for granted. The possible explanations for insufficient regional wage flexibility to maintain the labor market in equilibrium are related to shifts in labor demand and in labor supply. A lack of labor demand, productivity shortfalls, and a different industrial composition may limit regional employment. Similarly, labor supply, constrained by human capital and skills, demographic features (e.g., age, gender, and ethnicity), and a lack of mobility, may limit migration flows to expanding regions and increase regional unemployment. Both shifts in labor demand and shifts in labor supply, which interact and are not mutually exclusive, cause differentials in regional unemployment, making these differentials symptoms of inefficiencies in labor market adjustments.

Understanding the causes of high unemployment in particular areas and the nature of observed regional disparities in unemployment rates is an important step toward identifying the appropriate policy response. In the voluminous literature on this subject, two strands of thoughts and policies stand out: people-based policies and place-based policies. The first camp assumes geographical mobility which leads to a more even geographical distribution of employment and to the convergence of areas with low employment levels, while the second camp assumes that geographical context really matters and focuses on the issue of knowledge in labor policy intervention (Barca et al. 2012). On the one hand, people-based policies may generate shifts either in labor demand or in labor supply. For example, differences in the rate of unemployment may result from the weakness of the monopoly power of trade unions; grants that allow redundant workers to retrain in relevant skills; various government measures introduced to help school-leavers develop skills and job experience for the first time; special measures to encourage the long-term unemployed back into the labor force, which are policies aimed at the supply of labor; or through investment subsidies, tax breaks, and decrease of interest rates, which are policies aimed at the demand for labor (Begg et al. 2008). On the other hand, differences in unemployment rates may arise in reaction to place-based policies aimed at helping economically disadvantaged regions through subsidies, grants, or tax incentives. Hence, both people- and place-based policies could address high unemployment levels and high spatial disparities in unemployment rates, and it is not entirely clear which

type of policies will, in the long run, have more beneficial effects in tackling the local unemployment problems which have become so pervasive across many parts of the world.

References

- Barca F, McCann P, Rodríguez-Pose A (2012) The case for regional development intervention: place-based versus place-neutral approaches. *J Reg Sci* 52(1):134–152
- Barnes SA, Green A, Orton M, Bimrose J (2005) Redressing gender inequality in employment: the national and sub-regional policy ‘fit’. *Local Econ* 20(2):154–167
- Becker GS (1962) Investment in human capital – a theoretical analysis. *J Polit Econ* 70(1):9–49
- Begg D, Fischer S, Dornbusch R (2008) Economics, 9th edn. McGraw-Hill, London
- Blanchard O, Katz L (1992) Regional evolutions. *Brook Pap Econ Activity* 1:1–75
- Cheshire PC (1973) Regional unemployment differences in Great Britain. National Institute of Economic and Social Research. Cambridge University Press, Cambridge
- Coulson NE, Fisher LM (2002) Tenure choice and labour market outcomes. *Hous Stud* 17(1):35–49
- Courchene TJ (1970) Interprovincial migration and economic adjustment. *Can J Econ* 3(4):550–576
- Fields GS (1979) Place-to-place migration: some new evidence. *Rev Econ Stat* 61(1):21–32
- Glaeser E, Gottlieb J (2008) The economics of place-making policies. *Brook Pap Econ Activity* 2:155–239
- Goss E, Paul C (1990) The impact of unemployment insurance benefits on the probability of migration of the unemployed. *J Reg Sci* 30(3):349–358
- Hughes G, McCormick B (1987) Housing markets, unemployment and labour market flexibility in the UK. *Eur Econ Rev* 31(3):615–645
- Johnson GE, Layard PRG (1986) The natural rate of unemployment: explanation and policy. In: Ashenfelter OC, Layard R (eds) *Handbook of labor economics*, vol 3. North-Holland, Amsterdam, pp 921–999
- Marston ST (1985) Two views of the geographic distribution of unemployment. *Q J Econ* 100(1):57–79
- Martin R (1997) Regional unemployment disparities and their dynamics. *Reg Stud* 31(3):237–252
- McCormick B (1983) Housing and unemployment in Great Britain. *Oxf Econ Pap* 35(S):283–305
- McGee R (1985) State unemployment rates: what explains the differences? *Q Rev* 10(1):28–35
- Metcalf D (1975) Urban unemployment in England. *Econ J* 85(339):578–589
- Moffitt R, Nicholson W (1982) The effect of unemployment insurance on unemployment: the case of federal supplemental benefits. *Rev Econ Stat* 64(1):1–11
- OECD (2005) Employment outlook. OECD, Paris
- Oswald A (1997) The missing piece of the unemployment puzzle. Inaugural Lecture. University of Warwick, Nov 1997
- Overman HG, Puga D (2002) Unemployment clusters across Europe’s regions and countries. *Econ Pol* 34(17):115–147
- Partridge M (2001) Exploring the Canadian-U.S. Unemployment and nonemployment rate gaps: are there lessons for both countries? *J Reg Sci* 41(4):701–734
- Partridge M, Rickman D (1997) The dispersion of US state unemployment rates: the role of market and non-market equilibrium factors. *Reg Stud* 31(6):593–606
- Pehkonen J, Tervo H (1998) Persistence and turnover in regional unemployment disparities. *Reg Stud* 32(5):445–458
- Pissarides CA, Wadsworth J (1989) Unemployment and the inter-regional mobility of labour. *Econ J* 99(397):739–755
- Rodríguez-Pose A (2002) The European Union: economy, society and polity. Oxford University Press, Oxford

- Shaw RP (1986) Fiscal versus traditional market variables in Canadian migration. *J Polit Econ* 94(3):648–666
- Tatsiramos K (2009) Geographic labour mobility and unemployment insurance in Europe. *J Popul Econ* 22(2):267–283
- Taylor J, Bradley S (1983) Spatial variations in the unemployment rate: a case study of North-West England. *Reg Stud* 17(2):113–124
- Taylor J, Bradley S (1997) Unemployment in Europe: a comparative analysis of regional disparities in Germany, Italy and the UK. *Kyklos* 50(2):221–245
- Thirlwall AP (1966) Regional unemployment as a cyclical phenomenon. *Scott J Polit Econ* 13(2):205–219
- Winer SL, Gauthier D (1982) Internal migration and fiscal structure. Study prepared for the Economic Council of Canada. Supply and Services Canada, Ottawa

Dionysia Lambiri and Antonios Rovolis

Contents

8.1	Introduction	126
8.2	What Exactly Is Real Estate?	126
8.3	Developers and the Development Process	128
8.4	A Property and Asset Market Model	129
8.5	The “Internationalization” of Property and Asset Markets	134
8.6	The Housing Market	135
8.6.1	Demand, Supply, and House Price Determination	136
8.6.2	The Effect of Planning Controls on the Housing Market	138
8.6.3	Other Types of Intervention in the Housing Market	138
8.6.4	Housing Market and the Macroeconomy	139
8.7	Concluding Remarks	143
	References	144

Abstract

This chapter presents a comprehensive review of the fundamental concepts regarding real estate and housing markets. It aims firstly to provide an overview of the specific features of real property in general and housing in particular that make property a unique and multidimensional “good.” Building upon that, the chapter presents the key analytical tools extensively used in the relevant literature to capture the functioning of the real estate market as a set of interconnected markets, namely, the user (or space) market, the capital (or investment) market,

D. Lambiri (✉)

Geography and Environment, University of Southampton, Highfield, Southampton, UK
e-mail: d.lambiri@soton.ac.uk

A. Rovolis

Department of Economic and Regional Development, Panteion University of Athens, Kallithea, Greece
e-mail: rovolis@panteion.gr

and the development market. In this context, property development is examined as a process serving to reconcile long-run demand and supply imbalances generated in the user and investor markets. With regard to the housing market, after an overview of the key determinants of housing demand and supply, this chapter places its focus on the link between housing and the macroeconomy. Finally, the chapter explores the role of financial internationalization in the operation of real property markets and housing in particular, in the context of an increasingly globalized economy.

8.1 Introduction

In economic theory, land is ascribed a very important role as one of the main factors of production. However, real property and the operation of real estate markets have been relatively understudied in mainstream economics, despite the fact that its operation has important implications, not just for the efficiency of individual firms but also for the economy in aggregate. The value of new construction of buildings, either private or public, represents a significant component of the annual gross domestic product (GDP) in most countries. The value of existing buildings is the largest part of a nations' stock of wealth and represents one of the most important assets in the balance sheets of most firms. As such, the analysis of the built environment has increasingly become an important part of the curriculum in urban economics and in subdisciplines of housing and real estate economics, as well as related disciplines such as economic geography and political science.

A common perception is that real estate economics study the business and institutional dimension of property markets, whereas housing economics is primarily focused on public policy (e.g., for a more detailed analysis of the different approaches, see Arnott and McMillen 2006, pp. 142–144). This chapter attempts to present aspects from both these approaches and offers a comprehensive review of the basic concepts regarding real estate and housing markets.

8.2 What Exactly Is Real Estate?

Real estate analysis usually focuses on a specific type of property – for instance, housing or commercial properties. Actually, such an empirical analysis has a spatial dimension, that is, housing markets in a specific town or area. Property types are usually classified, for analytical and practical reasons, as housing and commercial properties; the latter category is in turn broken down to retail properties, hotels, offices, and industrial properties. Sometimes logistics is added as a separate subcategory of commercial properties; moreover, vacant land is also part of real estate. Investors in advanced real estate markets belong to two major categories: individual investors and institutional investors. Most textbooks analyze the economic behavior of the latter category, “not because there aren’t sizable individual investors engaged in real estate investing, but because they often team up with

institutions, which have the capital and set most of the ground rules for the investment program" (McMahan 2006, p. 59). The major institutional investors are life insurance companies, real estate investment trusts (usually referred as REITs), pension funds, open-end (they are called open-end funds because investors can periodically withdraw) and closed-end funds, and individual and institutional foreign investors.

Why is there a need for a separate treatment of built property from economics and other disciplines, when there is a vast body of existing literature regarding other types of goods? The answer is that real estate property in general, and housing in particular, has some distinctive features that make it unique.

Real estate property has primarily a *physical dimension*, as it involves a physical asset. At the same time, it has a *legal dimension* which refers to the property rights on the physical asset. In fact, what is traded in the property market is not the physical units of land and buildings but rather the legal rights or interests which exist over them. An important characteristic of this "physical" aspect of real estate which separates property from many other types of goods is its *durability*. The stock of real property constructed at any point in time lasts for many years, and its value can depreciate and/or become obsolete. The problem of property depreciation and obsolescence creates the need for maintenance which adds extra costs. It has to be noted that the depreciation or obsolescence of real estate property has to be construed in its economic and not "physical" meaning. It is possible that the "physical life" of a building may be longer than its "economic life" if nobody wants to use it either for housing or commercial use. Durability also implies that the supply for real estate property is considered "inelastic" – as property stock cannot easily be increased, at least in the short and medium run; as Baum (2009) observes, it is even more difficult to vary downward.

Another characteristic of real estate property is that it is highly *illiquid*. This is another way of saying that it is (very) costly to trade real estate property, as it involves direct cost such as taxes, legal fees, valuation fees, brokers' fees, and indirect costs. This latter category of costs is of great significance, as it includes the "information costs" regarding property, the risk associated with the property transaction, and other transaction costs (e.g., for a more extensive analysis of these costs, see Baum 2009).

Due to the aforementioned characteristics of heterogeneity, spatial fixity, high transaction costs, and asymmetric information among the market "players," it has often been argued that the real estate market is inefficient. The concept of what "efficiency" actually means for such a complex market as real estate has been inadequately theorized in the relevant literature. However, the judgment of inefficiency arises by reference to an "ideal" concept of efficiency, one which assumes a perfectly competitive market in equilibrium, characterized by a homogeneous product and rational, perfectly informed actors. The relevant research literature has primarily focused on the issue of *information efficiency* in the real estate market, not only in terms of how easy it is for participants to access all the potentially available knowledge before entering the market but also with regard to whether all relevant information is effectively capitalized into market prices.

Institutional approaches to the issue have “relaxed” the strict notion of efficiency and have offered a more pragmatic conceptualization, in which the degree of efficiency of the market in question is evaluated in comparison to its best potential (rather than absolute optimum) outcome. These approaches take into consideration the institutional environment with all the constraints that it imposes on the efficient operation of individual markets (see Evans (2004) for a detailed account of the issue of efficiency in real estate).

Due to availability of a wide variety of capital sources for both direct and indirect investment, the real estate market is increasingly becoming more “liquid” and less costly to trade and manage. Yet there remain significant “inefficiencies” in real estate, in both its user (or space) market and its investment market. High vacancy rates in commercial property, continuous fluctuations, and divergence across spatial submarkets, all of which are characteristics of real estate, are far from an economist’s vision of a perfect market.

8.3 Developers and the Development Process

Real estate developers can be classified into three categories. The first category is private-sector developers, which is the typical kind of developers. Another category is not-for-profit developers; they typically complete projects such as schools and hospitals. A third category, akin to the not-for-profit one, is public-sector developers. Private-sector developers charge development fees in order to cover their administrative costs and living expenses; they additionally earn reversionary profits on the sales of developed properties. Not-for-profit developers also earn developer’s fees, but they do not get any reversionary profits. This is also the case for public-sector developers; this category of developers is working in special markets, such as hospitals, schools, and governments buildings (Peca 2009).

Developers of all kinds have as their potential tasks to estimate future demand for the specific project in hand, and to calculate the costs, to obtain the necessary planning permissions, to find the necessary financial resources for the completion of the development project, to complete the construction phase, and to manage the constructed property (the developer can sell part of the property). A detailed description of these tasks can be found in Harvey and Jowsey (2004).

McDonald and McMillen (2006) define the stages of land development as follows: the first stage (initial contact by land broker) includes the site inspection, a preliminary market study and cost estimates, and an option contract with land owner; the second stage (option period) includes soil studies and engineering, feasibility appraisal and design strategy, finance plans, etc.; in the third stage (development period), the land is purchased, and loans have been secured; in the fourth stage (sales period), developers implement marketing programs, design controls, and facility management.

One potential breakdown of the development process is greenfield development, brownfield development, and greyfield development. The first category refers to development that takes place in “empty space,” for instance, farmland or forests.

Brownfield development is construction activity typically occurring in urban areas facing environmental degradation, for instance, in defunct industrial plants. The last category (which is rather controversial as a term) refers to cases in which existing “underutilized” buildings are improved (“redeveloped”).

Developers engage in one or more activities (or phases) of the development process. The development process is broken down to the site acquisition, cost planning, market planning, financial leasing, project timing and scheduling, property management, approval of the plan, architecture design, engineering design, and actual construction. Other suggested development “phases” in the relevant literature include concept and initial consideration, site appraisal and feasibility study, detailed design and evaluation, contract and construction, and marketing, management, and disposal (see Ratcliffe et al. 2004). What becomes evident from the complex process of development is that a real estate project involves a great array of professions, organized by the “developer” (for a more extensive analysis, see Peca 2009).

8.4 A Property and Asset Market Model

The most celebrated model of real estate markets is DiPasquale and Wheaton (1996). A similar model was presented by Fisher (1992). It is the staple model in advanced undergraduate and postgraduate courses in urban economics and real estate markets and in popular textbooks and papers alike (e.g., Achour-Fischer 1999; Geltner et al. 2007). Their model is elegant in its simplicity because it introduces two basic markets for real estate property: the market for “use” of built stock (“Property Market: Rent Determination,” in the model’s parlance) and the market of property as an asset (called “Asset Market: Valuation”). This analytical “device” connects these two different “functions” of real estate property and has a pedagogical significance of its own; nonetheless, it has the power to describe the “dynamics.” Yet Colwell (2002, p. 24) argued that the initial model reveals very little about short-run adjustments of real estate and construction markets in the context of a “comparative static analysis” (also see DiPasquale and Wheaton 1996, p. 11).

We present an abridged version of the DiPasquale and Wheaton model (D-W model). The focus on this model is due to its explanatory potential for the basic mechanics of real estate and construction markets, and secondly because some of its most important points go amiss while trying to grasp the complexities within a real economy. Of all the different aspects of the real estate markets assessed in the model, this section concentrates on the structure of the various markets, their interconnection, and on the repercussions an exogenous change would have on these markets. The model comprises four separate diagrams combined in one, thus forming a “cross”-type diagram, which allows to study the way in which changes in one market may affect others, as well as the feedback mechanisms built in the model.

The first market analyzed is that of “use” of built space. The fact that it is always the first to be analyzed does not imply that it is more significant than the other markets. Indeed, the most important feature of this model is the “interaction” of the

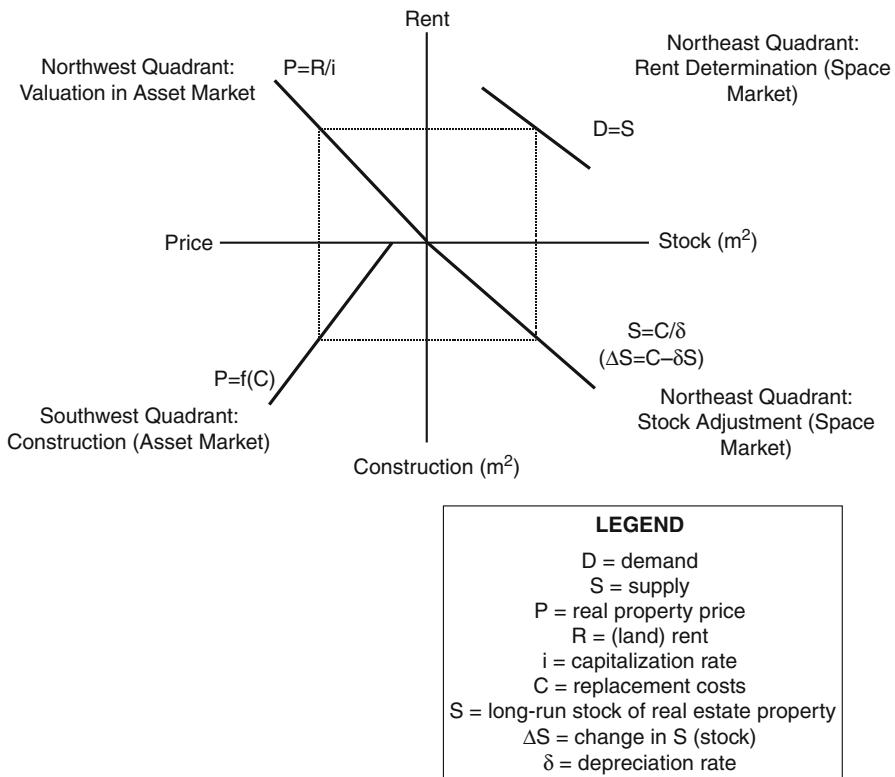


Fig. 8.1 The DiPasquale and Wheaton model of property and asset markets

markets. The diagram of this property market presents the physical stock of built property described as a vertical inelastic supply curve (the “quantity” of this supply and demand diagram, usually expressed in square feet or meters) on the horizontal axis, and on the vertical one the “rent” determined by space use (the “price” in the diagram); thus, at the point of equilibrium, supply equals demand. Demand is a function of the level of rent and the state of the general economy (DiPasquale and Wheaton 1996, p. 8). In some books, the four different quadrants which constitute the “cross” diagram are called northeast, northwest, southwest, and southeast, respectively, following an anticlockwise turn (this convention is followed here). Thus, the “property market for use” is the northeast of Fig. 8.1.

The second, northwest, part of Fig. 8.1 describes the valuation of property in asset markets; here, current rent is related to real property prices (needless to say that in this northwest part of the diagram, the horizontal axis, which is the price of real estate property, increases from right to left). The important question is how rent is “transformed” (related) to prices; this happens via the simple equation $P = R/i$, where P is real property price, R is (land) rent, and i a capitalization rate. This equation begs the next question which is exactly what a capitalization rate is.

The previous equation can be rearranged to $i = R/P$, but, again, this is rather tautological.

Capitalization rate is exogenously determined by a number of factors, such as forecasts about the demand and supply of space markets (e.g., Geltner et al. 2007, p. 24). Note that the capitalization rate is not the nominal interest rate, as it can be construed as a real rate of risk (Colwell 2002, p. 26). If the line in this quadrant is closer to (or further from) the rent axis, that is, this line is “steeper” (“flatter”), it then follows that a given level of rent will be “translated” to “lower” (“higher”) prices for property. A number of factors can make this line “steeper,” for instance, an increase in long-term real interest rates (or the inflation rate), an increase in taxes on real estate property, or greater perceived risk for real estate. To summarize, a “steeper” line in this quadrant is a graphic expression of an increase of the capitalization rate; such an increase will lower real estate prices.

The third part of the diagram, that is, the southwest quadrant, is the one describing the operation of the construction – the development industry. The line in this quadrant is the short-run supply curve of the construction industry. Here, the construction investment is gross, as it contains both the new investment on built space and the replacement of depreciated buildings. The construction line is a short-run curve, which in this diagram does not emanate from the intersection of axes as it is supposed that a minimum level of price is needed in order to have construction activity. This activity is linked with the equation $P = f(C)$, where P is real property price and C are the replacement costs. Factors that increase the construction cost will shift the line away from the construction level axis; such factors are, among others, an increase of short-run interest rates, stricter building or zoning regulations, etc. Colwell has argued that it is possible to reduce the range of prices over which there is no construction activity (Colwell 2002, p. 27). However, such an alteration would compromise the look of the diagram. Geltner et al. have observed that the slope of the construction line depicted in this quadrant represents long-run costs in the supply of the built space (Geltner et al. 2007, p. 27).

The last market of the diagram is that of the southeast quadrant, which represents the way in which the flow of new construction is converted into property stock. This relationship is described by the equation $S = C/\delta$, where S is the long-run stock of real estate property, C are replacement costs, and δ is the depreciation rate. Depreciation in this context means that older buildings are “either abandoned and demolished or converted to other uses” (Geltner et al. 2007, p. 27). The slope of the line in this quadrant represents the “speed” of depreciation process, that is, a steeper line represents faster depreciation. Note that in some cases this line can look as one half of a single line, if the other half is the line of the northwest quadrant. Of course, this is not the case, as the two lines are completely different.

It is argued above that this diagram can be used to show fluctuations and long-run equilibrium in real estate markets. Due to space limitations, only one exogenous change in the system is presented – a demand shift in the northeast quadrant; the interested reader can explore more exogenous changes in the three basic references of this section, that is, Colwell (2002), DiPasquale and Wheaton (1996), and Geltner et al. (2007). This case is presented in Fig. 8.2. The inner box depicts the initial state of

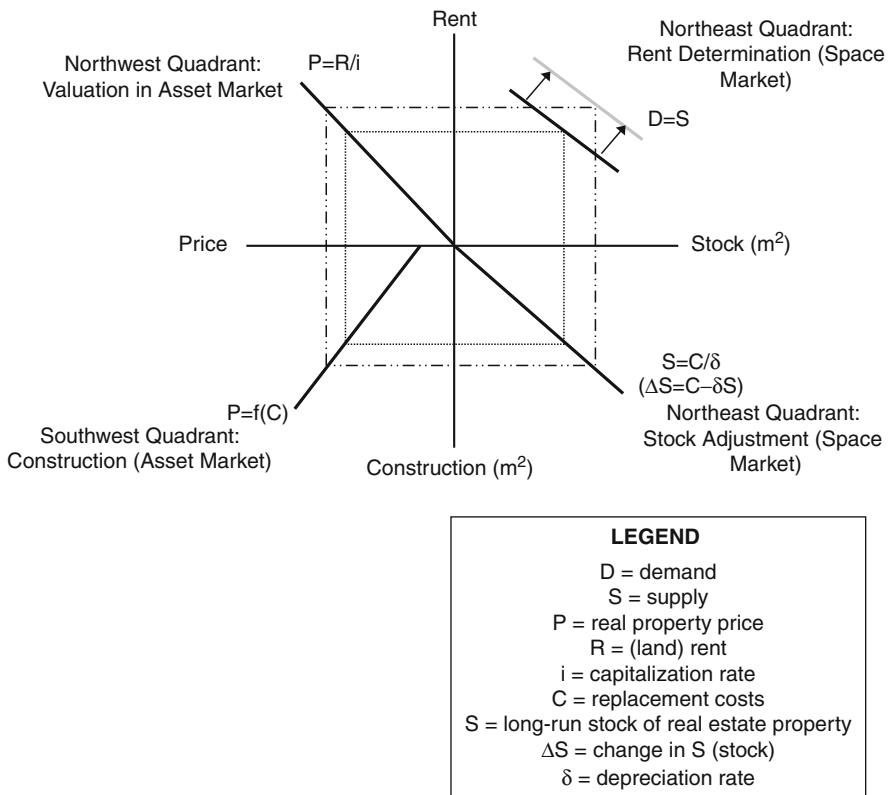


Fig. 8.2 The DiPasquale and Wheaton model: The effect of demand growth on property and asset markets

equilibrium and the outer box the new equilibrium achieved after the exogenous shock that is the increase of the demand for the use of built space. What is missing is the “path” of the adjustment process between the initial and the new equilibrium.

Figure 8.3 illustrates the adjustment process (this diagram is originally presented in Colwell 2002; a more extensive analysis is offered there). The demand for built space use shifts toward the right, and in the new temporary equilibrium, the supply is still the same (as there is not enough time for supply to adjust), and the rent is higher. This leads to an increase of property price in the northwest quadrant and, in turn, in an increase of the gross property construction in the southwest quadrant. This translates as an increase of the real property stock, via the southeast quadrant. As a result, in the northeast quadrant, there will be a shift of the supply curve to the right, which will eventually result in a decreased property rent. This process will continue until there is equilibrium in all four markets. An interested reader can find an altered D-W model in Colwell (2002), where the vertical supply line of northeast quadrant has been replaced by a long-run supply curve with negative slope.

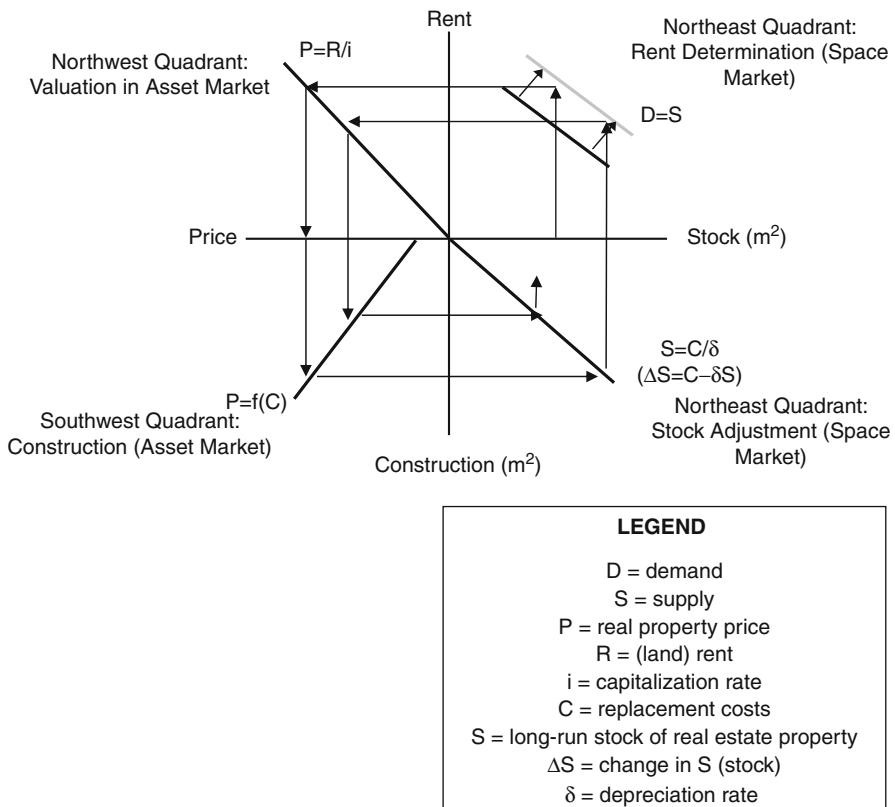


Fig. 8.3 The DiPasquale and Wheaton model: The adjustment process after demand growth

(In this case, there is an extensive presentation of the adjustment process toward the new long-run equilibrium). This equilibrium is presented in [Fig. 8.2](#).

In the D-W model, real estate property, in its “asset incarnation” (i.e., in the northwest quadrant), is compared to alternative investment options, such as bonds and securities. One of the real estate characteristics is its “lumpiness” (Baum 2009), as property comes in large and uneven sizes, meaning that it cannot be diversified in the way other kinds of assets can. Thus, the creation of a portfolio based on real estate property can be rather challenging.

The overview of the complex characteristics of real estate markets shows that it is important to incorporate in any analysis both built space for “use” and real estate property as an asset. Formal analytical models make evident that real estate markets reach a state of equilibrium after a prolonged process; they also show that one of the most important dimensions of built property analysis is real estate finance. Real estate finance globally links the local and specific markets of real estate properties. This trend intensified in the second half of the twentieth century with increasing internationalization of the estate markets.

8.5 The “Internationalization” of Property and Asset Markets

The DiPasquale and Wheaton model presented above analyzes the interconnection between the different markets related to real estate property. It also shows, however, that one of the most important dimensions of the analysis is real estate finance. Modern real estate markets are based on underlying physical assets, which are the basis of a whole financial structure. One crucial difference that separates real estate market from other markets, for example, manufacturing, is that large real estate investors can own directly the underlying physical assets. How then is it possible for small investors to own these physical assets that are worth large sums? The answer is the use of liquid-traded property vehicles. These vehicles include real estate investment trusts (REITs), which own directly real estate property. Since they have access to capital, they can also build additional properties (Block 2006, p. 8). Other ways small investors can own physical assets include other unlisted funds, property derivatives, and mortgage-backed securities (Baum 2009).

It is argued earlier on that real estate markets have essentially a local “nature.” In one sense, the seller of a particular property is a monopolist, as that property is the only one in the specific floor, at the specific side, at the specific street. There is, however, the argument that there is an increasing internationalization/globalization of the real estate markets (e.g., Seabrooke et al. 2004). This trend has started, or at least intensified, in the second half of the twentieth century – when the confluence of several historical events, economic processes, and technological changes transformed the financial markets crucially linked to real estate investment and construction activity.

The seed for this “internationalization” of financial markets was sowed with the breakdown of the Bretton Woods agreement and the shift to floating exchange rates in the early 1970s. This led to the increase of the use of existing financial products and the creation of new ones. This “internationalization” hinged on the existence of new technologies such as computers, and sophisticated telecommunications, which make possible the (instantaneous in most cases) international transfer of vast amounts of money, ending to a large degree in the real estate markets. The question, however, is where these vast amounts were coming from. The brief answer is that there were several sources of global capital accumulation in the last 50 years (see Goldberg (2004) for more details). The oil crises of the 1970s and the trade surpluses of the oil-producing countries created the “petrodollar” markets. Other countries, for instance, Japan in the 1980s and recently China, had also accumulated huge amounts of capital resulting from trade surpluses; the economic growth of Asia led to groups such as the Southeast Chinese to amass a lot of capital. In Europe, the process toward a single economic market and the European Union has also had a deep impact on the real estate markets (Nicholls 2004 offers an extensive analysis). In the last decades, differences in exchange rates between the dollar and other currencies have created a flow of capital mostly from Europe to the United States. The overall conclusion of these fundamental changes in the real estate markets is that at one level, they keep their local characteristics, but at another level, associated with the financial (asset) “nature” of property markets, real estate is increasingly “internationalized.”

8.6 The Housing Market

Housing constitutes a major sector in national economies, and therefore, the way housing markets function is closely interlinked with many elements of the macroeconomy. The functioning of housing markets can be explored through the standard tools of supply and demand analysis; however, housing, like all real estate, is a multidimensional commodity with inherent complexities and distinctive characteristics. This means that the analysis of housing markets needs to take into account these complexities.

First of all, like all the other forms of property, housing is a *heterogeneous* good. Each house is made up of a bundle of characteristics – both structural (number of rooms, garden size, presence of a garage, etc.) and location specific (proximity to a park, coastal or mountain location, proximity to schools, etc.). These characteristics affect how much people are willing to pay for a particular housing unit (for a detailed account of how the heterogeneous nature of housing is captured and theorized in house price models, please refer to ► Chap. 9, “[Housing Choice, Residential Mobility, and Hedonic Approaches](#)” in this handbook).

Secondly, housing is spatially fixed. In terms of the analysis of the housing markets, this characteristic gives rise to two analytical issues:

- (a) Since housing cannot move, households need to move to homes. This means that the availability or absence of appropriate housing in a location will influence location decisions of households and therefore generate mobility. This is one way in which housing is linked with the macroeconomy and the labor market in particular. Moreover, the spatial fixity of housing is the reason why housing development gives rise to major planning debates, as there can be mismatches between demand and supply of housing in different locations.
- (b) The housing market is not an “aggregate entity” but a series of interconnected location-specific submarkets: these can be international, regional, or urban. Analyses at these different spatial scales give rise to different analytical and policy issues. For instance, while the study of international housing markets focuses primarily on the global interaction of housing markets, the study of regional housing markets highlights regional differences and views housing markets as one factor reflecting (as well as explaining) regional growth variations. In the case of urban housing markets, the focus is on intraurban location decisions of households, spatial segregation, and social exclusion in the housing market. A key difference between international/national housing market analysis and regional/urban analysis is that in the former cases, the spatial dimension is absent.

Finally, like all the other forms of property, housing is a *durable* good. Housing stock lasts for many years (60–100 years), which implies that the existing stock is quite substantial in relation to the flow of newly constructed housing. Most importantly, the characteristic of durability automatically implies that other than a “user function,” housing has an investment aspect as well. Therefore, when looking at the motives for holding residential property, analyses should consider and distinguish between demand for housing for consumption (occupation) and demand for housing

for investment purposes. However, housing other than a durable “consumer” good and an investment good is also a *social good*: governments are committed to provide decent accommodation to disadvantaged parts of the population. Housing subsidies and tax incentives exist in order to make housing more affordable. These can be argued to generate distortions in the functioning of the market and the price of housing.

All these complex characteristics of housing, combined with the aforementioned high transaction costs that generally apply to the property market (legal fees, estate agent fees, search costs), constitute sources of inefficiency in the functioning of the housing market. Traditional neoclassical space/access models do not take into account these complexities. As Maclennan and Tu (1996) note: “*Although the insights of the standard neo-classical model are very important, it is clear that economic phenomena such as mis-information, commodity variety, space, time and the nature of the market itself are all victims of the abstraction process*” (p. 388). In these models, housing is approached as a one-dimensional homogeneous product, and the focus lies predominantly on the relationship of housing with the wider urban structure rather than on the operation of housing markets. In these models, the price of housing is demand determined, as supply of housing seems to respond “unproblematically” to changes in demand conditions. As such, these models do not take into account factors that affect the supply of housing, such as future expectations for house price changes, changes in construction costs, time lags, uncertainty, and speculation in the land and housing markets.

8.6.1 Demand, Supply, and House Price Determination

One of the building blocks of housing economics is the assumption that in the short-run, housing supply is inelastic. Factors such as uncertainty, monopolistic ownership, high transaction costs of property changing hands, and, most importantly, the time new housing takes to be constructed imply that short-run supply is measured in terms of just the existing housing stock, which is assumed to remain unchanged. Therefore, in the short run, house prices are assumed to be dependent only on changes in housing demand.

The factors that determine housing demand (H^d) can be expressed as

$$H^d = f(\underbrace{Y, PH, PC, r, Cr, W, T}_{\text{Short Run Determinants}}, \underbrace{\text{Pop}, HR, MG}_{\text{Long Run Determinants}}) \quad (8.1)$$

where Y denotes income, PH house price, PC price of other goods, r mortgage interest rates, Cr credit availability, W wealth, T housing taxation, Pop population size, HR headship rates, and MG migration.

Housing demand in the longer run is determined by (a) natural population growth and (b) population growth due to migration and (c) changes in headship rates. Fluctuations in real incomes and interest rates are considered the most

important determinants of housing demand especially in the short run. As a result, they constitute key causes of house price fluctuations. However, the extent to which house prices “respond” to changes in real incomes depends on the income elasticity of housing demand. Different studies have provided different estimates of income elasticities, ranging from values of as high as two to close to zero. Observed differences in income elasticities can result from the use of different measures of income: for instance, income elasticities are higher when permanent (predicted) income rather than current income. Moreover, the use of aggregate rather than individual data would inflate observed income elasticities. For instance, Mayo (1981) suggests that most US estimates find income elasticities greater than one when aggregate data are used and below one when individual data are used. Finally, income elasticities will vary significantly for rental as opposed to owner-occupied housing (Meen 2001). The estimation of income elasticities is a core issue in housing studies, as it plays an important role with respect to both urban structures and housing policies. Indicatively, Meen (2001) highlights the importance of the study of income elasticities, by pointing out that estimated income elasticity greater than one might imply that income increases would lead households to relocate away from the city center to the suburbs.

In the long run, housing supply is responsive to changes in demand for housing. New housing development takes place in the longer run, and the housing supply curve is no longer inelastic due to the addition of newly constructed units in the existing housing stock.

Housing supply (H^S) at any period can therefore be expressed as

$$H^S = (1 - \delta)H^S_{-1} + Q \quad (8.2)$$

where H^S_{-1} denotes the existing housing stock (stock “inherited” from previous period), δ depreciation rate, and Q new housing construction.

This expression shows that the total housing stock for a city/metropolitan area at a specific point in time equals to housing constructed in (“inherited” from) previous periods, after taking into account the rate of depreciation, plus the new supply of housing due to construction in the current period. In turn, this new supply Q , which affects H^S in the longer run, can be expressed as a function of

$$Q = h(PH, CC, r) \quad (8.3)$$

where CC denotes construction costs including the cost of land for development.

New housing construction is negatively affected by increases in interest rates and construction costs, whereas it is positively related to changes in house prices. The sensitivity of housing construction to interest rate changes illustrates how monetary policy can have an effect on housing construction cycles. In the long run, housing supply is considered to be more responsive to changes in house prices and therefore not perfectly inelastic. Excessive price decreases or increases experienced in the short run will “ease,” and, in the long run, prices will move in line with construction costs.

Differences in house prices in different locations (as well as variations in price among different housing types) will reflect consumer preferences. Willingness to pay sends “signals” to the market and will affect the price the construction industry is prepared to pay for acquiring land for housing developments of certain types and in certain locations.

8.6.2 The Effect of Planning Controls on the Housing Market

The responsiveness of housing supply is affected by the availability of developable land, which in turn depends on the land use regulations in place. Studies of the European (including the UK) and US housing markets demonstrate a strong positive correlation between the level of strictness of land use regulations, low house supply elasticities, and house prices. Land use and planning policies can impose restrictions on land availability and new construction and affect – directly and indirectly – the profitability of land developed for housing purposes. Moreover, such policies change the spatial benefits of particular locations through infrastructure provision. Monk et al. (1991) offer a systematic approach of the complex relationship between planning and the ways it can affect land and house prices. He describes that an effective planning system directly affects the supply of land made available for housing not only by restricting the total quantity of land for development but also by restricting the location of the land that is made available. This pattern restricts the way in which land is developed and determines the timing of development. Moreover, the existence of housing submarkets and the heterogeneity of housing as a complex good means that development constraints in one location cannot be “counterbalanced” by increased availability of land for housing development in another location. The degree of substitutability of house types and locations will determine whether constraints applying for a certain type or location will affect another.

Development control and planning regulations do not only affect the supply of housing and the amount of land put forward for housing development purposes; they can also affect housing demand: negatively by the imposition of costs and by long delays or uncertainties associated with obtaining planning permissions, and positively by setting environmental and design guidelines that will affect potential buyers’ willingness to pay for housing and consequently increase the value of housing as an investment good. In this latter case, planning restrictions are viewed as a mechanism that can correct negative externalities such as overcrowding and incompatible land uses (nevertheless, density controls could be considered to have negative effect for those households demanding small plots at high densities, such as low-income households).

8.6.3 Other Types of Intervention in the Housing Market

Other than planning restrictions, housing markets in most countries feature various degrees of government involvement, which can take several forms, from direct

subsidies (housing allowances, public housing) and fiscal interventions (taxation, mortgage incentives) to market regulations such as rent controls or tenure protection legislation (For an example of the complexities of modelling housing choice in the context of regulated urban housing markets, see Aufhauser et al. (1986)). The reasons for an active role by the government are multifaceted; one reason for government intervention is the well-documented inefficient functioning of the housing market which makes government action necessary in order to achieve Pareto efficiency; moreover, there is a “welfare” aspect based on the premise that housing is a necessary good and society ought to provide housing if an individual cannot afford it; therefore, supporting housing consumption and investment serves as a mechanism to redistribute income and wealth. However, what needs to be noted here is that relevant research suggests that, while housing market regulation such as public housing provision is undoubtedly a means to alleviate affordability problems, it can be far from unproblematic, due to its high supply costs, as well as due to the fact that can often exacerbate (or even be the cause of) socioeconomic segregation.

Whitehead (1999) notes that housing policy has been approached – and analyzed – very differently in the European and American literature. Traditionally, in the American literature, local and national governments are seen as actors that can influence private investment in housing through housing taxation, rent controls, subsidies, and zoning. These forms of regulation are largely local policies and, as such, can be evaluated through cross area comparisons. In Europe, on the other hand, housing regulation is approached by the relevant literature mainly in relation to welfare and social policy. Supply subsidies and rent controls are primarily decided at the national level while often mediated by local governments. Consequently, such policies apply either at a national level or across population groups/households.

8.6.4 Housing Market and the Macroeconomy

Housing constitutes a major sector in national economies and represents the largest share of household assets. As a result, changes in house prices will have an effect on many aspects of national and regional economies and, most importantly, incomes, consumption, and the labor market.

There is the argument that one of the most important aspects of housing is its connection to the welfare state; this line of reasoning originates in the work of Kemeny (1980) and Castles (1998). Schwartz and Seabrooke (2009) have provided a classification of western countries according to the characteristics of their housing markets. One category refers to the countries in which the housing market is highly commodified. These countries have, according to Schwartz and Seabrooke's (2009) terminology, a *liberal market* in which houses are treated as assets, and where there are high owner-occupation rates and high property tax revenues, and where mortgages as percentage of GDP are high. These markets are more integrated in the global financial system. Another type is *corporatist markets*, which have lower

owner-occupation rates, but high percent of mortgages versus GDP. In both of these housing market categories, there is strong market stratification. The *statist-developmental* category refers to countries in which housing is regarded as a social right. Here, the owner-occupation rate is relatively low, as are also mortgages as percentage of GDP. In this category of countries, property tax revenues are low. There is, finally, the *familial* category in which housing is a familial social good. It is characterized by high owner-occupation rates and low percentage of mortgage on GDP (see Schwartz and Seabrooke's (2009) extensive description).

These differences across western countries imply that the housing market will affect, and will be affected by, the overall economy in different ways. In countries with a liberal economic system, it is expected that households which invested in residential property would accumulate wealth over their life cycle. These households would prefer (and, thus, vote for) lower taxation and lower interest rates. In these societies, pensions, *ceteris paribus*, would be expected to be low and houses to be a substitute for retirement income (for a detailed analysis, see Schwartz and Seabrooke 2009).

The strong links of the housing market and the macroeconomy and its effects on both aggregate demand and supply have been widely discussed (Meen 2003; Case et al. 2005). The relevant literature specifically focuses on the effects of changes in house prices on consumer expenditure (Muellbauer 1990; Holmes 1993; Parkinson et al. 2009), the links between the housing and the labor market (Cameron and Muellbauer 1998; Cannari et al. 2000; Engelhardt 2003; Henley 1998), and the analysis of housing booms and busts in the context of increasingly globalizing economies. The following sections look at these issues in turn.

8.6.4.1 House Prices and Consumer Expenditure

The relationship between house prices and consumer spending is not only a causal one: factors such as changes in interest rates, or people's expectations about future incomes, affect both the demand for consumer goods and services and the demand for housing and therefore have an effect on house prices. Especially with regard to future expected income, empirical evidence suggests that this affects the behavior both of homeowners and renters and therefore plays a significant role in the co-movement of consumer spending and house prices. However, theoretical and empirical research in the field also discusses the causal relationship as well, as house price fluctuations are expected to contribute to fluctuations in consumer spending. There are various ways in which this link can be explained.

The degree of credit market liberalization plays a big role in the way house price growth affects growth in consumer spending, since it leads to house price increases being translated into increased borrowing for consumption purposes. The expansion of mortgage markets in recent years has strengthened the linkages between house prices and consumption. Moreover, the liberalization of credit markets has increased the sensitivity of housing markets (and consumption) to interest rate changes. Housing as an asset can be made more "liquid," either through equity release schemes or through remortgaging. Credit products that are secured against

the value of a house allow consumers to access the equity that their house “contains” without having to sell it.

In the context of the UK housing market, Muellbauer (1990) examines this causal relationship between house price growth and excessive growth in consumer expenditure in the UK. In a similar context, Holmes (1993) discusses the phenomenon of equity withdrawal (the borrowing on mortgage more than it is required to finance the purchase of a house) in the UK and provides evidence of the positive impacts from equity withdrawal on the average propensity to consume. Boelhouwer (2000) reports similar findings for the case of the Netherlands, where the dramatic increase of house prices in the 1990s led to a high rise in consumer spending, partly financed by the overvaluation of the Dutch homes. Again, in the Dutch case, credit market liberalization and the opportunity of commercial banks to expand their offer or mortgages led to increase borrowing, of which only a small part went to the purchase of new homes. Parkinson et al. (2009) examine comparatively the phenomenon of mortgage equity withdrawal in the cases of Britain and Australia, concluding that equity borrowing “*is not just about using housing wealth as routinely as an ATM; rather housing wealth is funding some substantial ‘one-off’ or sustained expenditures*” (p. 385). Their research suggests that in the recent financial crisis, housing wealth took more of an insurance role (a safety net) rather than a pure consumption one, where households used it to meet expenditure demands associated with job loss, childcare, and general welfare.

In the case of the USA, macroeconomic studies find evidence that there is an important relationship between the wealth effect associated with housing and the propensity to consume. Indicatively, the study of Case et al. (2005) uses a panel of quarterly data for US states and a panel of annual observations on 14 developed countries to find a large and significant effect of housing wealth upon household consumption – while at the same time evidence on the wealth effect of financial assets seems weak.

8.6.4.2 The Effect of Housing on the Labor Market

Households seldom make employment decisions independently from housing decisions. The functioning of local and national housing markets will affect labor mobility through house price levels and housing tenure structure. High relative house prices in a region can be a factor that discourages in-migration (net of amenities), and relative house prices can have an effect on choosing where to migrate. Moreover, as banks and mortgage lenders allocate mortgages based on loan-to-house value and loan-to-income ratios, first-time buyers interested in buying property in more prosperous regions could face cash-flow problems if mortgage rates increase, while, conversely, residents in richer regions could be in an advantaged position of being able to use their equity in order to reduce borrowing or to move up the property ladder by moving to other regions (Cameron and Muellbauer 1998).

Tenure structures in the housing market are also related to labor mobility. The relevant literature suggests that even though in more recent years, local authority tenants have become more mobile, they are still the least mobile, followed by owner occupiers, whereas private tenants have the highest levels of mobility.

Even though mobility between regions can be the result of reasons irrelevant to employment choice, there is strong international evidence to support the link between the housing market and labor mobility. Indicatively, in the case of the UK, Henley (1998) finds evidence to suggest that negative equity in the early 1990s impaired households to sell their property and move and therefore reduced their ability to find a better job match elsewhere. Labor immobility leading to inability to match vacancies and consequently causing labor market inefficiencies is also noted in Cameron and Muellbauer (1998) who also find that, as levels of owner-occupation rose in the UK, the influence of relative house prices on net migration rates has also risen. In the case of Italy, Cannari et al. (2000) use market price data at a provincial level to examine whether the housing market played a role in the decline in internal migration between 1965 and 1995. They provide evidence to suggest that differences in the cost of housing between the north and the south of Italy have restrained migration flows and are an important factor in explaining falling patterns of labor mobility.

8.6.4.3 Housing Markets in the Global Economy

In recent years, the integration of increasingly deregulated international financial markets and the general business cycle linkages has led to the co-movement in house prices across countries globally. Strong similarities in house price fluctuations across different countries are not a characteristic of just the recent house price boom: studies of house price data for the period 1970–1992 from OECD countries show that house price dynamics are interdependent, even though there is no evidence supporting the existence of an international house price cycle (Englund and Ioannides 1997). In more recent years, the degree of international housing market synchronization has been widely researched (Kim and Renaud 2009). Even though individual countries might exhibit differences in levels of house price growth, there are significant ties between housing markets of big European countries and the USA. Studies have indicated significant correlations between real house price trends between the USA and the EU, with EU countries following US house price trends with an approximate 2-year time lag.

The recent economic crisis and the international slump in housing markets starting in 2007 have stimulated further research of the interdependences of international housing markets. The housing market has played a key role in the economic boom of recent years, as well as and the subsequent downturn. This can be explained by the fact that, as housing and housing-related spending (for renovation, maintenance, as well as furniture and appliance purchases) “pull” large amounts of capital, they allow high levels of equity extraction, which can generate large amounts of household debt. The aforementioned increasing liberalization of housing markets, with the confluence of the persistent (and unsustainable) increase of house prices and low interest rates, encouraged less affluent homebuyers to enter the US housing market, triggering the subprime mortgage crisis. Banks and mortgage lenders sold mortgage derivatives (“packaged” mortgages) to institutional buyers, the value of which soon “crashed” as some households started defaulting on their mortgage payments. The consequence of uncontrolled mortgage

lending and loan distribution was disastrous not only for major US financial institutions but for institutions worldwide, some of which had to be either bailed out by their national governments or were taken over by other financial institutions (for a more extensive analysis of the main reasons for the collapse of the housing market, as well as of the role of new “exotic” financial products in the crisis, see Schwartz 2009).

In recent years, comparative studies on international house price trends have also looked at the issue of housing affordability. Apart from the destabilizing effect that the recent international house price boom has had on national economies, it also raised major concerns with regard to housing affordability (measured as the house price to earnings ratio). On the one hand, factors mentioned earlier, such as low interest rates and favorable borrowing and mortgage conditions designed to enhance affordability to potential homebuyers by reducing monthly payments, made the housing market more “accessible” to lower income and younger households (first-time buyers). However, in most countries, wages and household incomes of middle- and low-income households did not rise proportionately to rising house prices, leading to ownership affordability problems. Kim and Renaud (2009) review relevant studies for different countries (France, Spain, New Zealand, Australia) which provide evidence to suggest that low interest rates and the subsequent rise in house prices caused declines in homeownership rates.

8.7 Concluding Remarks

The analytical intricacies arising from specific features of real property in general and housing in particular make the study of the operation of the real estate market a complex task. This chapter has touched upon these complexities and has highlighted the multidimensional nature of the property market as a set of interconnected markets – the user market, the investment market, and the development market, all of which are invariably linked to the macroeconomy.

On the one hand, demand for commercial property is derived directly from economic activity; land and property are very important inputs in the production process. As a result, in most developed economies, the decline in manufacturing and the growth of the service sector were reflected in decreased demand for industrial property and increased demand for retail and office space. In fact, more recently, the fast pace of globalization in information and communications technology (ICT) and the emergence of a plethora of dot.com firms in the 1990s (all of which were seen as the foundations of a new so-called “weightless” economy), gave rise to a debate as to whether the role of real estate as an asset in the production process would weaken. This raised the following question: will demand for real estate decline, as information technology starts substituting real estate in production functions? Evidence suggests that the relevant debate of the “death of real estate” has been largely exaggerated. Naturally, developments in ICT have brought significant changes in business processes, which have affected traditional sectors and have caused radical changes in all sectors of commercial real estate (for an extensive analysis, see Dixon et al. 2005).

However, real estate space is increasingly adapting, leading to the emergence of new types of real estate, and the reconfiguration of existing real estate spaces that take into account changes in customer and service provider relationships.

With regard to the housing market, this chapter has underlined its strong interaction with the macroeconomy and in particular with household incomes, consumer spending, wage formation, unemployment, and migration. Construction of housing (and commercial property alike) is highly dependent on monetary policies through the effects of interest rate fluctuations on construction costs. The effects of the recent housing boom and subsequent bust have been a clear illustration of this interconnectedness, in the context of an increasingly globalized economy and in times of financial liberalization and internationalization.

References

- Achour-Fischer D (1999) An integrated property market model: a pedagogical tool. *J Real Estate Practice Educ* 2(1):33–43
- Arnott RJ, McMillen DP (2006) A companion to urban economics. Blackwell, Oxford
- Aufhauser E, Fischer MM, Schönhof H (1986) A disaggregated probabilistic approach to a regulated housing market with an emphasis on the demand side: the Vienna case. *Papers Regional Sci Assoc* 60:133–153
- Baum A (2009) Commercial real estate investment: a strategic approach, 2nd edn. EG Books, London
- Block RL (2006) Investing in REITs, 2nd edn. Bloomberg Press, New York
- Boelhouwer PJ (2000) Development of house prices in the Netherlands: an international perspective. *J Hous Built Environ* 15(1):11–28
- Cameron G, Muellbauer J (1998) The housing market and regional commuting and migration choices. *Scott J Polit Econ* 45(4):420–446
- Cannari L, Nucci F, Sestito P (2000) Geographic labour mobility and the cost of housing: evidence from Italy. *Appl Econom* 32(14):1899–1906
- Case KE, Quigley J, Shiller R (2005) Comparing wealth effects: the stock market versus the housing market. *Adv Macroeconom* 5(1):1–32, Article 1
- Castles FG (1998) The really big trade-off: home ownership and the welfare state in the new world and the old. *Acta Politica* 33(1):5–19
- Colwell PF (2002) Tweaking the DiPasquale-wheaton model. *J Hous Econ* 11(1):24–39
- DiPasquale D, Wheaton WC (1996) Urban economics and real estate markets. Prentice Hall, Englewood Cliffs
- Dixon T, Thomson B, McAllister P, Marston A, Snow J (2005) Real estate and the new economy: the impact of information and communications technology, Real estate issues series (RICS foundation). Blackwell/Oxford, Malden/UK
- Engelhardt GV (2003) Nominal loss aversion, housing equity constraints, and household mobility: evidence from the United States. *Journal of Urban Economics* 53(1):171–195
- Englund P, Ioannides YM (1997) House price dynamics: an international empirical perspective. *J Hous Econ* 6(2):119–136
- Evans AW (2004) Economics, real estate and supply of land, Real Estate Issues Series (RICS Foundation). Blackwell/Oxford, Malden/UK
- Fisher JD (1992) Integrating research on markets for space and capital. *Real Estate Econ* 20(2):161–180
- Geltner DM, Miller NG, Clayton J, Eichholtz P (2007) Commercial real estate analysis and investments, 2nd edn. Thomson South-Western Publishing, Mason

- Goldberg MA (2004) Local property markets and effective flexible market institutions. In: Seabrooke W, Kent P, Hong How HH (eds) International real estate: an institutional approach, Real estate issues series (RICS foundation). Blackwell/Oxford, Malden/UK, pp 96–129
- Harvey J, Jowsey E (2004) Urban land economics, 6th edn. Palgrave Macmillan, New York
- Henley A (1998) Residential mobility, housing equity and the labour market. *Econ J* 108(447):414–427
- Holmes MJ (1993) Housing equity withdrawal and the average propensity to consume. *Appl Econ* 25(10):1315–1322
- Kemeny J (1980) Home ownership and privatization. *Int J Urban Regional Res* 4(3):372–388
- Kim KH, Renaud B (2009) The global house price boom and its unwinding: an analysis and a commentary. *Hous Stud* 24(1):7–24
- MacLennan D, Tu (1996) Economic perspectives on the structure of local housing systems. *Hous Stud* 11(3):387–406
- Mayo SK (1981) Theory and estimation in the economics of housing demand. *J Urban Econ* 10(1):95–116
- McDonald JF, McMillen DP (2006) Urban economics and real estate: theory and policy. Blackwell, Malden
- McMahan J (2006) The handbook of commercial real estate investing. McGraw-Hill, New York
- Meen G (2001) Modelling spatial housing markets: theory analysis and policy, Advances in urban and regional economics series. Kluwer, Boston
- Meen G (2003) Housing, random walks, complexity and the macroeconomy. In: O'Sullivan T, Gibb K (eds) Housing economics and public policy, Real estate issues series (RICS foundation). Blackwell/Oxford, Malden/UK
- Monk S, Pearce B, Whitehead C (1991) Planning, land and house prices: a literature reviewmonograph 21, Property research unit, Department of land economy. University of Cambridge, Cambridge
- Muellbauer J (1990) The housing market and the U.K. economy: problems and opportunities. In: Ermisch J (ed) Housing and the national economy. Avebury, Aldershot
- Nicholls DC (2004) Emerging institutions in Europe. In: Seabrooke W, Kent P, Hong How HH (eds) International real estate: an institutional approach, Real estate issues series (RICS foundation). Blackwell/Oxford, Malden/UK, pp 155–172
- Parkinson S, Searle BA, Smith JS, Stoakes A, Wood G (2009) Mortgage equity withdrawal in Australia and Britain: towards a wealth-fare state? *Int J Hous Policy* 9(4):365–389
- Peca SP (2009) Real estate development and investment: a comprehensive approach. Wiley, Hoboken
- Ratcliffe J, Stubbs M, Shepherd M (2004) Urban planning and real estate development, 2nd edn. Spon Press, London
- Schwartz HM (2009) Origins and consequences of the U.S. subprime crisis. In: Schwartz HM, Seabrooke L (eds) The politics of housing booms and busts. Palgrave Macmillan, New York, pp 188–207
- Schwartz HM, Seabrooke L (2009) Varieties of residential capitalism in the international political economy. In: Schwartz HM, Seabrooke L (eds) The politics of housing booms and busts. Palgrave Macmillan, New York, pp 1–27
- Seabrooke W, Kent P, Hong How HH (2004) International real estate: an institutional approach, Real estate issues series (RICS foundation). Blackwell/Oxford, Malden/UK
- Whitehead CME (1999) Urban housing markets: theory and policy. In: Cheshire P, Mills ES (eds) Handbook of regional and urban economics, vol 3. Elsevier, Amsterdam, pp 1159–1594

Housing Choice, Residential Mobility, and Hedonic Approaches

9

David M. Brasington

Contents

9.1 Introduction	148
9.2 Residential Mobility	148
9.3 House Price Hedonics	155
9.4 Conclusions	162
References	164

Abstract

This chapter explores the literature on residential mobility and house price hedonics. Residential mobility studies the decision of economic agents to move or not and, if they move, their choice of new residence. Topics covered in this chapter include the theory behind the move-or-stay decision, modeling intra- and interregional moves, empirically validated determinants of moving, and macro- and microlevel studies on mobility. Next, house price hedonics explain the price of a house as the sum of all the things that give a house value, from structural characteristics like the number of full bathrooms to public services and neighborhood characteristics that the house experiences. The chapter discusses the theory behind hedonics, applications of the technique, and empirical approaches to identify hedonic house price studies and second-stage hedonic regressions of the demand and supply of characteristics that give a house its value.

D.M. Brasington

Department of Economics, University of Cincinnati, Cincinnati, OH, USA

e-mail: David.Brasington@UC.Edu

9.1 Introduction

Everyone moves at some point in their life, and if they are lucky, they move into a dwelling. This chapter takes up the topics of residential mobility and house price hedonics: the movement of people from one dwelling to another and the factors that influence the price of the dwelling they move into.

The first part of this chapter covers residential mobility. First, the theoretical underpinnings of residential mobility are covered: why do people move? Life cycle models, cost-benefit models, and neighborhood change models are discussed. The chapter then discusses attempts to model moves within a region and between different regions. The factors influencing moves are explored, including the role of job changes, social capital, public policy, climate, and differences in the mobility of men and women. It includes a look at macrolevel and microlevel studies. The chapter finishes with a discussion of whether jobs follow people or if people follow jobs.

The second part of the chapter covers house price hedonics. First, the theoretical framework of the hedonic method is discussed. The issue of estimating a second-stage demand or supply regression from the first-stage hedonic is then explored, with particular attention to identifying the second stage from the first. The various applications of the hedonic method are discussed: (i) capitalization of taxes and public services into house price, (ii) quantifying the relative importance of non-market goods to house prices, (iii) constructing hedonic price indexes, (iv) evaluating policy alternatives, (v) applications to real estate finance, and (vi) testing for market segmentation. The discussion of hedonics concludes with an exploration of a handful of ways researchers have tried to identify hedonic parameters, including instrumental variables, spatial econometrics, the borders approach, and the mixed index approach. The final part of the chapter offers concluding thoughts and directions for future research.

9.2 Residential Mobility

Residential mobility studies the decision of economic agents to move or not and, if they move, their choice of new residence. Residential mobility is a key driver of economic development. By means of agglomeration economies, residents make a city more productive the larger a city becomes. The forced relocation of peasants to large urban areas in the early years of the Soviet Union was an attempt to create the conditions for a revolution of the proletariat described by Marx but was also an attempt to kick-start economic development. The migration of rural poor to urban areas of China has been a driver and consequence of the economic development of China's cities in our day.

Most of the literature in regional science on residential mobility takes an existing urban setting as its starting point. Winstanley, Thorns, and Perkins (2002) provide

a valuable background for the literature, noting that mobility of households within an urban area was traditionally considered a negative phenomenon because it interfered with social networks and moves were often provoked by deteriorating housing stock or neighborhoods or by adverse health events in a family. When Rossi (1955) challenged this view, a new literature was born on the premise that it is perfectly normal for a household to change locations over time. As people marry, have children, age, and experience other *household-formation events* like divorce and retirement, their housing needs change and they move to meet these new needs. These needs include house-specific needs like extra bedrooms but also encompass the need for schools, shopping, leisure and employment opportunities, and proximity to family and friends.

Winstanley, Thorns, and Perkins (2002) divide the literature on residential mobility into three strands: *life cycle models*, cost-benefit models, and *neighborhood change models*. From a regional science point of view, all models can be reduced to cost-benefit models if the costs and benefits are defined appropriately. Recent work by Rabe and Taylor (2010) illustrates the linkage between life cycle models and neighborhood change models, finding that life cycle events may be associated with a move to a better neighborhood (having a baby) or a worse one (a single person moving out of a parent's basement or a husband's unemployment).

Brueckner (2011) provides a useful overview of residential mobility using Tiebout (1956) sorting within a single urban area as a starting point. Tiebout sorting says that under certain conditions, households vote with their feet to end up in a municipality whose levels of public services exactly match consumer preferences, without the community changing its public service level.

In theory, with a head tax, an urban area consists of a collection of municipalities with internally homogeneous public service level demand. If income levels drive public service demand, municipalities are stratified by income. In the real world, *Tiebout sorting* is imperfect: there are a limited number of municipalities to choose from, so sorting is restricted to the “closest” match, leaving internally heterogeneous communities. Urban areas characterized by city-county governments are especially restricted in opportunities for Tiebout sorting, while there are better opportunities for sorting in large urban areas with many municipalities.

But even in a theoretical world with unlimited potential for Tiebout sorting, *economies of scale* provide a limit to the number of municipalities that will form. In the presence of important economies of scale, a minimum number of residents are required to lower the average cost of public service provision to a level competitive with its neighbors. That is, even if community A provides exactly the level of public service desired by a household, the household may choose community B if the cost of public service provision, and therefore taxes, is significantly lower than A, while the level of public service provision is sufficiently close to the most-preferred level. Scale economies may occur at different population levels for different services, which may result in a municipality

providing its own police services but cooperating with other municipalities for schooling services.

If higher income levels equate to higher preference for public service levels, a rich municipality chooses a high public service provision with a high head tax. But if models allow preferences to deviate from income levels, lower-income individuals may wish to sort into a high-service municipality. A sufficiently high head tax may be enough to discourage high-preference poor people from entering the rich municipality. Whether because of income levels, a prohibitive head tax, or both, a poor community forms with a low head tax and a low level of public service.

Property taxes finance a large share of local public services in the USA, and to the extent that they do, it allows more internal heterogeneity within a municipality. All else equal, a poor person would like to consume the larger amount of public service. But the taxes required are too high under a head tax. If a property tax replaces the head tax and the poor person can find a small house in the rich municipality, the poor person can pay a small property tax and still enjoy the large amount of public service.

The influx of poor residents into a rich municipality will have fiscal consequences for the rich residents. Because the poor residents pay less for the service, the municipality may run a fiscal deficit on these residents, a deficit that must be made up for by rich residents if the most-preferred level of public services is to be maintained. In response, rich residents may choose to engage in fiscal zoning, so that new construction pays for the cost of the public services provided to the new residence.

The literature long held that job changes primarily affect moves between urban areas but that households do not generally move within an urban area in response to a change in jobs. However, Clark and Davies-Withers (1999) show that, in fact, a person who takes a new job in the same urban area as the old job is 2.4 times more likely to find a new residence than people who experience no change in job. Such *intra-urban moves* are more likely for young households, single-earner households, and renters; in contrast, older households, dual-income households, and homeowners are more likely to commute to their new job from their current residence.

Apart from mobility within an urban area, households also may move from one urban area to another. Such long-distance moves to new labor markets are often spurred on by changes in employment opportunities, including households moving after retirement (Clark and van Lierop 1986). Changes in human capital are often cited as a cause of *interregional migration*.

While changes in employment opportunities are important to mobility, recent research explores reasons why such moves do not always happen. For instance, Kan (2007) explores the role of social capital in the decision to move, finding evidence that the more social ties a household has in an area, the less likely it is to move to a new location within the same labor market and that social ties make a move to a new labor market even more unlikely.

The issue of household mobility, be it intraurban or interregional, involves three decisions: (1) whether to move, (2) where to move, and (3) what to move into.

These decisions may be viewed as sequential as in a nested framework (e.g., Fischer and Aufhauser, 1988) or as a simultaneous set of decisions. The decision of whether to move has been modeled as *housing dissatisfaction* (e.g., Hanushek and Quigley, 1978). Before Hanushek and Quigley, most research on residential mobility was descriptive. Hanushek and Quigley motivate their work by noting that moving is a dynamic response to a household's changed circumstances. The decision to move depends not only on housing demand but also on search costs, the transaction costs of moving, and the distribution of house prices available to a prospective mover. Hanushek and Quigley model the decision to move as a probit function where the dependent variable is the difference between a household's housing consumption and its desired housing consumption in the next period, divided by its desired housing consumption in the next period. They use a data set of renters from Phoenix and Pittsburgh who were interviewed initially, then one year later, then a year later again. Housing demand is estimated by an ordinary least squares model that regresses monthly rent on income, wealth, age, household size, household education level, race, and whether the apartment has a refrigerator, air conditioner, or stove. The percentage difference between housing consumption expenditures and predicted housing consumption expenditures is Hanushek and Quigley's measure of disequilibrium. The decision to move is modeled as a function of demographic characteristics like age, income, race, and household size. The authors extend their model to a multinomial probit case of whether a household searches and moves, searches for new housing without moving, and does not search for new housing. Hanushek and Quigley's model is readily extended from housing dissatisfaction to neighborhood dissatisfaction, providing a link between housing dissatisfaction and neighborhood change models.

More recent research on household mobility has focused on the role of public policy. California's Proposition 13 prohibits reassessment of the base-year value of a property except in the case of new construction or a change in ownership. The law provides incentive for residents to stay in their current dwelling. Recent work by Ferreira (2010) estimates the impact on household mobility by analyzing recent amendments to Proposition 13 that allow residents over 55 years of age to sell their house and still transfer the tax savings to a new house. Ferreira finds that 55-year-old homeowners are 25 % more likely to move than comparable 54-year-olds.

A wide variety of approaches to modeling mobility have been used in the literature, as discussed by Clark and Van Lierop (1986). These include linear programming models, gravity models, discrete choice, and behavioral models. The methods are described below, and readers may refer to Clark and Van Lierop (1986) for specific examples of research using each method.

Perhaps the oldest approach to modeling residential location is the linear programming model, which is the counterpart to the Alonso location model. In the linear programming model, households choose a market basket of non-housing goods to maximize the ability of the household to spend on housing. The outcome of choosing the optimal basket results in the household choosing the optimal housing location, optimal in the Pareto sense that no household can relocate to

increase its expenditures on rent without displacing another household, thus reducing the other household's expenditures on rent.

Gravity models were sometimes used to model residential location instead of linear programming models. *Gravity models* are derived from Newton's law of gravity and are also known as distance decay models, *spatial interaction models*, and *origin-destination flow models*. These models describe the attraction between two objects, in this case people in one location and their attraction to another location. It is common in the residential mobility literature to parameterize spatial interaction models with a log-linear functional form to look at the aggregate movement of households between areas. These log-linear models regress the number of moves between two areas as a function of explanatory variables such as the net migration from each area, the control variables, the distance between each area, and the distribution of flows between areas. Shortcomings of these models, particularly their failure to consider spatial dependence, are discussed in LeSage and Fischer (2010).

The micro counterpart to the log-linear spatial interaction models is the use of discrete choice models, pioneered by Hanushek and Quigley (1978). Even today, most research follows this methodology. Households receive utility from choosing a dwelling, and the utility they receive affects the probability that they will select the dwelling. The discrete choice framework is used to model the move-or-stay decision, the rent or own decision, as well as the decision to select a new place of residence from a variety of locations. The discrete choice framework is also used in more complex procedures like sequential choices of move versus stay, then choice of a specific neighborhood from a finite option set, then a choice of a specific dwelling within the neighborhood.

Behavioral models of residential mobility examine a household's level stress, disequilibrium, or, as in Hanushek and Quigley (1978), dissatisfaction with housing consumption (discussed earlier in this chapter). Much attention has been given to the search process involved in relocation. Often, such models contain a stopping parameter that guides when a household finalizes its decision, such as when the expected gap between the expected utility resulting from additional search and the highest utility available is small enough. *Search models* have incorporated preferences over housing consumption and neighborhood consumption, as well as incorporating the preferences of single-earner households, dual-earner households, and the preferences of households with children that allow the children's preferences to enter into the decision calculus.

Perhaps due to data constraints, many of the early studies on residential mobility were macro-oriented, dealing with the moves of an aggregated number of people between geographic areas like urban areas, census tracts, or municipalities. The increased availability of data has led to the micro approach – the examination of the residential mobility decisions of a single household – gathering the lion's share of attention in the literature. Nevertheless, recent research reviewed by Dieleman (2001) has rediscovered the macro side of the equation, uncovering new factors important for mobility between geographic areas. *Turnover rates of the housing stock* and housing prices vary across regions and across time. The renewal of the

housing stock allows for more potential choice of housing by area residents, while low housing prices make home ownership more affordable to young people and household formation by young people into renter-occupied dwellings more feasible. Household mobility is twice as prevalent in the American South as in the American Northwest. Research suggests that moves are more common where the local economy is stronger: such areas are marked by increased construction of new dwellings and an influx of young workers. Endogenous growth in the number of young workers also plays a role, as Southerners are more likely to raise children and Northwesterners are more likely to raise dogs. Differences in public policy might also help explain variations between regions in mobility, with the urban growth boundaries of Portland providing a marked contrast with the urban sprawl of Atlanta and the regulatory difficulty of constructing new dwellings in California contrasted with the relative lack of zoning laws in Texas. Internationally, such differences help explain the relative lack of housing stock turnover and household formation in Europe. There is some truth to the joke of the Italian man who proudly proclaims, “Before I was 35, I lived with my mother; now that I’m over 35, my *mother* lives with *me*.” Variation in *housing prices* tends to persist over time, although there are places where the variation in house prices over time is greater, such as the coastal regions of the USA, which experience boom and bust cycles to a greater degree than inland cities. Higher house prices are found in larger urban areas, more rapidly growing urban areas, and areas where construction costs – including regulatory approval costs – are higher. Direct government intervention in the housing market also affects the price of housing, such as rent controls, government construction of low-income dwellings, urban growth boundaries and development taxes, and government treatment of renting versus owning.

Mobility is a subject with many aspects to be investigated: renting versus owning, within versus interurban mobility, and mobility by education, marital status, child status, racial composition, and more. Perhaps surprisingly, little research has investigated the *mobility of women* relative to men. The field commonly assumed that women were less mobile than men without engaging in much empirical research to confirm or refute the supposition. Faggian, McCann, and Sheppard (2007) find evidence that women in the UK are more mobile than men. They point out that previous research on the issue almost entirely ignores the role of human capital in the relative migration of men and women. Faggian, McCann, and Sheppard study a sample of UK university graduates. They follow these individuals from their homes to university and from university to place of first employment. After controlling for human capital and regional economic conditions, they find women are more interregionally mobile than men, arguing that this increased mobility represents attempts by women to overcome gender bias in the labor market. Faggian, McCann, and Sheppard also find that racial minorities, older individuals, unemployment in the individual’s hometown, and distance of the individual’s hometown to London reduce mobility of the four types of migrants studied, relative to nonmigrants. They use multinomial logit, dichotomous logit, and conditional logit regression approaches.

In recent decades, the role of climate in attracting new residents has drawn a lot of research attention. The population shift from New England and the Midwest in the USA to the “Sunbelt” areas of the South and West is attributed in large part to climate. Rappaport (2007) challenges the claim that the invention of air conditioning is responsible for the shift, finding increased population not only in areas with warmer winters but also in areas with cooler, less humid summers. Rappaport also notes that the shift to nice weather is driven not just by elderly retirees but also is happening at almost the same rate by working-age people. Rappaport details the literature on *climate and residential mobility*, pointing out that the traditional approach is the *compensating differential approach*: estimating the wages that workers forego to live in places with a better climate. But Rappaport notes that there are several important drawbacks to the compensating wage differential approach used by previous studies. One such problem is that the level of detailed observations required is only available for places with over 100,000 people. This population requirement introduces sample selection bias because data is only available from places where a lot of people have already chosen to live. A second problem is the inability to properly control for individual-specific and house-specific characteristics. This is particularly a problem because high-income individuals may choose to live in areas with a high quality of life, so that the unobserved characteristics of these individuals are likely positively correlated with quality of life variables, including climate. Given the limitations of the compensating differential approach, Rappaport instead studies climate’s role in changes in quality of life and productivity. Weather’s contribution to changes in quality of life and changes in productivity has occurred for four reasons: (i) the decline of agricultural employment has made weather less important for agricultural productivity, (ii) air conditioning and improved heating technology have made extreme temperatures more bearable, (iii) rising incomes have increased the consumption of good weather, and (iv) the utility value of weather has increased because of the rise in the number of affluent retirees. His approach is unable to distinguish between the quality of life and the productivity effects of climate, and the approach cannot quantify the size of the effects, but it avoids the problems of the compensating differential approach and opens up a new angle of attack with which to research a long-standing issue. Rappaport uses county-level data, which provides a large number of observations and full coverage of the USA, avoiding the sample selection problem of using urban area-level data. The annual growth rate of population density is regressed as a function of average January high temperature, average July high temperature, precipitation, and other controls. Population growth is found to be statistically significantly related to high temperatures in January and July in most regressions.

Like Rappaport (2007), Partridge and Rickman (2003) also investigate a long-standing issue with a new approach. For decades, regional scientists have investigated whether *jobs follow people or people follow jobs*. The lack of finality in the issue stems from the fact that both job growth and population growth are endogenous, leading multiple researchers to call this a “chicken and egg” problem. Job growth draws new residents, while an influx of new residents creates new jobs.

Early attempts to identify the direction of causality employed simultaneously estimating employment and migration equations using instrumental variables, but such an approach suffers from the difficulty of finding appropriate instruments. It also fails to estimate short-run versus long-run responses. *Vector autoregression (VAR) models* have been used more recently to try to disentangle the job-people question, but previous attempts have fallen short in various ways. Some VAR studies assumed that all contemporaneous employment innovations were labor-demand shocks. Other VAR studies fail to include relevant equations in their models, such as migration equations or wage equations, making it impossible to disentangle labor demand and supply. Instead, Partridge and Rickman use a structural VAR model that incorporates a labor market model. The labor market model is used to create long-run identifying restrictions for the structural VAR model. Partridge and Rickman also distinguish between *labor-supply and labor-demand shocks* noting that wages respond in opposite directions to each type of shock. Finally, Partridge and Rickman decompose labor-supply shocks into those due to new residents and original residents. They use data from the 48 contiguous United States from 1969 to 1998, finding labor-demand shocks account for about 46 % of the variance in employment forecasts. Migration (labor-supply) shocks account for about 33 % of the variance, and internal labor-supply shocks the remaining 21 %. So while people seem to follow jobs more than jobs follow people (46 % vs. 33 %), when you sum the 33 % from migration and 21 % from internal labor-supply shocks, it is clear that the role of supply is as important as that of labor demand. Partridge and Rickman's model suggests that Sunbelt states are more influenced by migration shocks, and Rustbelt, Farm Belt, and Energy states are more influenced by labor-demand shocks.

9.3 House Price Hedonics

House price hedonic regressions attempt to break the price of a house into each aspect of the house that gives it value, from structural characteristics like house size to neighborhood characteristics like crime rates and to public finance characteristics like tax rates. Adding up the price of all the things that give a house its value will tell you the market value of the house. But a house is a bundle of services, each providing value but few being easily added onto or subtracted from a house. That is, while it may be possible to find the market value of adding another full bathroom to a house by getting a quote from a contractor, it is not as easy to take the same house and purchase an extra unit of air quality or public safety for it. The house price hedonic estimation is a way to use statistics to estimate the value that an extra unit of public safety would provide toward a house.

Residential mobility is linked to house price hedonics through many channels. As households *Tiebout sort* into an urban area, their demand for specific housing characteristics will be reflected in house price through the relative supply of that attribute in the area. If houses with high-quality public schools are in short supply relative to demand, households will have to pay a premium to live in such houses.

Housing turnover affects residential mobility, but the lack of new housing will tend to push the price of newer houses up, especially in the face of supply restrictions like development taxes and urban growth boundaries. And while residential mobility is likely to be less pronounced in urban areas with low *population growth rates*, the level of house prices is likely to be lower in such areas as well.

Although its use goes back to the early 1900s, the modern era of house price hedonics began when Sherwin Rosen (1974) detailed the theoretical underpinnings of the method. Rosen notes that the number of rooms in a house provides value to the house, but a house is a bundle of attributes, including the number of rooms. If we let z be the amount of housing consumed, then $z = (z_1, z_2, z_3, \dots, z_n)$ lets the amount of housing consumed be a function of the n characteristics of the house, so that summing z_1 through z_n yields the amount of housing consumption. These quantities can be converted to prices p so that $p(z) = p(z_1, z_2, z_3, \dots, z_n)$: the price of a house $p(z)$ is a function of the price of its attributes.

Buyers and sellers come together in the market to determine the price of the attributes of a house, the end result of which determines the selling price of the house. Consumers have a budget constraint in which their incomes are exhausted on purchases of housing attributes and a numeraire good. Consumers maximize utility over housing attributes and the numeraire good subject to a budget constraint. The first-order conditions of the constrained maximization equate the marginal rate of substitution for house characteristic z_i to p_i , the marginal price of characteristic z_i . And because the house is a bundle of characteristics that cannot be easily bought and sold separately on the market, each p_i is an *implicit price*, not an explicit price.

Consumers bid for housing attributes given a fixed level of income and utility. A bid is the marginal rate of substitution between a housing attribute z_i and money, which shows the consumer's reservation demand price for an additional unit of z_i . Next, $p(z)$, the minimum price a consumer must pay in the market for various quantities of z_i , is graphed along with consumers' bid functions, and tangencies show the optimal consumption of an attribute by a consumer.

Rosen next considers *producers of housing*, which are assumed to operate in a perfectly competitive environment. Producers maximize profits by producing additional housing attributes until marginal cost equals marginal revenue. A producer's offer function shows the unit prices a producer will accept for various house designs, given a fixed level of profits. Producers maximize their offer prices subject to the price they can get in the market, $p(z)$. Rosen graphs a producer's offer function for characteristic z_i along with the price $p(z)$ obtainable in the market, and tangencies between these curves reveal the optimal amount of characteristic z_i the firm should offer, given the optimal quantity of other housing characteristics.

Rosen next puts consumers' bids together with producers' offers, and tangencies denote a joint envelope all along $p(z)$, the implicit price function of characteristic z_i . Rosen notes that $p(z)$ may be nonlinear: the price of a characteristic may change as more of it is consumed.

Rosen's work would have been influential enough if he had stopped there, but Rosen goes on to describe how the supply and demand for *implicit house price*

hedonic characteristics may be estimated. Step 1: regress $p(z)$ as a function of all z_i . This is the house price hedonic. Step 2: compute the marginal implicit prices $\partial p(z)/\partial z_i = p_i(z)$ for each buyer and seller, evaluated at the actual quantities consumed. Step 3: use the marginal implicit prices endogenously in a system of supply and demand for characteristics, being careful to include some exogenous shift characteristics in each equation. Step 4: estimate supply and demand simultaneously, making sure both supply and demand equations are identified from each other.

Later, James Brown and Harvey Rosen (1982) claimed the need to modify Sherwin Rosen's method of estimating a second-stage supply and demand system for hedonic attributes. They point out that while Sherwin Rosen was careful to identify the supply curve from the demand curve and vice versa, neither the supply nor the demand is identified from the initial house price hedonic in Step 1 above. All the information from the house price hedonic is embedded in the marginal implicit prices included in the second-stage supply and demand regressions, so that attribute prices are not independent of errors in the second stage.

Because of the difficulties in *identifying the demand for a hedonic attribute* like air quality from the underlying house price hedonic, most researchers stop at estimating a first-stage hedonic regression. However, there have been suggestions for how to identify second-stage supply and demand equations from the first-stage hedonic. The first, and the most frequently applied, is Brown and Rosen's (1982) suggestion to segment the housing market.

Market segmentation is the identification strategy taken by Brasington and Hite (2005). Brasington and Hite regress the natural log of house prices on twenty-two explanatory variables, including the focus variable, DISTANCE TO HAZARD, representing the distance from each house to the nearest environmental hazard. They run separate first-stage hedonic regressions for the six largest urban areas in Ohio, using a *spatial Durbin model* to help control for the influence of omitted variables. They use the parameter estimate of DISTANCE TO HAZARD to calculate the marginal implicit price of *environmental quality* for each of the 44,255 houses in the sample, although values are averaged at the census block group level (5,051 census block groups). The implicit prices for the six urban areas are based on six different parameter estimates of DISTANCE TO HAZARD, but all implicit prices in the sample are pooled to estimate a single second-stage demand curve for environmental quality. The implicit prices are treated endogenously using instrumental variables. Brasington and Hite find a small, statistically significant relationship between house prices and environmental quality in five of the six urban areas. They estimate the demand for environmental quality using two-stage least squares, limited information maximum likelihood, a fixed effect model, and a spatial Durbin model. There seems to be no relationship between environmental quality and lot size in the spatial regression. However, house size seems to be a substitute for environmental quality, and school quality and environmental quality seem to be complements.

Other than Brown and Rosen's (1982) market segmentation by geography, researchers have used or promoted identification of demand and supply from

first-stage hedonics by segmentation by time, functional form assumptions, and instrumental variables approaches. Of particular note is a paper by Cheshire and Sheppard (1998), who use a spatial lag of the two nearest houses as an instrument for implicit prices. Cheshire and Sheppard obtain similar results for their instrumented and so-called underidentified demand system, suggesting that while identification of demand from initial hedonic regressions is theoretically important, it does not affect regression results much. Also of note are Ekeland, Heckman, and Nesheim (2002), which claim that Rosen (1974) was right. Specifically, they claim that when all the information in the hedonic model is exploited, second-stage demand equations are identified from the first-stage hedonic even in single markets and without having to impose an arbitrary functional form assumption. The problem with second-stage models is that the hedonic model is generically nonlinear, and it is only when people use a linearization that identification problems arise. They show theoretically that a normal linear quadratic model achieves identification, but they do not actually estimate such a model.

Far more common than two-stage demand models are models that simply apply the first-stage hedonic house price regression to various problems of interest to regional scientists. These applications fall into at least six categories: (i) capitalization of taxes and public services, (ii) quantifying the relative importance of non-market goods to house prices, (iii) constructing hedonic price indexes, (iv) evaluating policy alternatives, (v) applications to real estate finance, and (vi) testing for market segmentation.

One of the first *applications of the hedonic method* to regional science was to investigate the capitalization of taxes and public services into constant-quality house prices. The idea of capitalization of taxes, for example, is that, holding public services and everything else constant, low taxes provide value to a house. Low taxes provide value to a house in the current period, the next period, and all successive periods for the life of the house (or land) subject to a discount rate. The sum of this discounted stream of value boosts house price above what it otherwise would be, so that low taxes are “capitalized” into constant-quality house price. Oates (1969) is the most-cited paper in the field of public finance capitalization, although Orr (1968) preceded that article or was at least contemporaneous. Oates found capitalization of both property taxes and school spending and claimed that it proved Tiebout (1956) was right. But the interpretation of *capitalization* has a large literature of its own. Several papers claim that capitalization should not occur in theory and that Tiebout’s paper itself would not predict capitalization. Other papers say that capitalization does exist and that its existence is a result either of a scarcity of desirable public service/tax combinations, local governments using more resources than necessary to produce public services, or local governments spending too much or too little on public services. Another paper says that both major views of capitalization are right: it should occur and should not occur simultaneously, depending on whether the house is in an area of high or low housing supply elasticity. See Brasington (2002) for a detailed discussion and empirical exploration of this issue.

The house price hedonic has been used extensively to quantify the relative importance of *non-market goods*. The difficulty for economists and regional scientists in studying nonmarket goods is that there is no observable market price. Some researchers estimate the market price by contingent value surveys or estimating travel cost models; others exploit the housing market. If air quality imparts value to a house, it should be possible to see how much value it imparts to a house by regressing house price as a function of house characteristics, neighborhood characteristics, and air quality. The partial derivative of house price with respect to air quality yields its marginal implicit price, from which a marginal willingness to pay can be calculated. Other nonmarket goods whose value has been estimated by house price hedonic regressions include public school quality, crime rates, and proximity to lakes, parks, hospitals, churches, shopping malls, and airports.

Hedonic price indexes are used to track how the price of a typical house changes over time. The house price hedonic is first estimated. A researcher then plugs in the sample means of the explanatory variables which, together with the parameter estimates for each variable, yield the typical house's estimated price in the base year (index = 100). The regression is run again for new time periods, with researchers plugging in the *old* sample means for the explanatory variables and using the new parameter estimates to come up with a new value for the index. There are different ways to construct hedonic price indexes. Gatzlaff and Ling (1994) compare indexes constructed from traditional hedonics based on sale price, hedonics based on assessed value, and the repeat sales method. The repeat sales method follows the sale price of the same houses over time. In theory, it provides an unbiased price index because, unlike the traditional house price hedonic, there are no unobserved characteristics to bias the price – as long as the quality of the house remains constant between sales. The criticisms typically levied against the *repeat sales method* are that (i) it suffers from sample selection bias, as certain starter homes and undesirable homes may sell more often than other houses; (ii) that the same houses may have been remodeled between sales, so that researchers are not really comparing the same house over time; and (iii) that the repeat sales method wastes a tremendous amount of data, because only a small subset of houses sell more than once over a small time period. These and other criticisms of the repeat sales model are detailed by Haurin and Hendershott (1991). Gatzlaff and Ling get around these criticisms by having information about houses that did not sell during the time period, using the entire set of houses for the city of Miami between 1971 and 1991, and controlling for houses that have been substantially remodeled between sales. Gatzlaff and Ling find that all methods provide precise measurements of house price indexes, both in levels and in changes between years, although there was more noise in quarterly price changes. It is especially noteworthy that indexes based on assessed value were similar to those based on actual sale prices, because assessed property value data is much easier and cheaper to obtain than sale price data.

The house price hedonic can be used to *evaluate policy alternatives*. This can be done in a cross-sectional study or a study measuring changes over time. For a cross-sectional study, a researcher pools data from a group of houses that does

not have policy x and a group that does. A dummy variable for policy x is included in the regression, and a statistically significant parameter estimate for the dummy variable tells whether the policy has a positive, negative, or no effect on house prices. Care must be taken to control for every possible influence, so that omitted variables are not driving the dummy variable's parameter estimate. The same procedure can be done across time, with the sample containing house sales before and after the policy change and the dummy variable reflecting sales after the policy change.

There are numerous applications of the house price hedonic to real estate. Researchers have used the technique to estimate the discount or premium for having an unusual house for the neighborhood, a motivated seller discount, differences in sale price for houses transacting with a real estate broker or for-sale-by-owner, and the effect on price of selling by way of a sheriff's auction or bank sale.

As mentioned before, the two-stage hedonic demand technique described by Rosen (1974) is most often conducted by using segmented housing markets in the first-stage regression. Another use of the hedonic technique is to test for *market segmentation*. A different house price hedonic regression is run for every area suspected of being a distinct housing market. If the parameter estimates are significantly different from each other, the houses are deemed to come from segmented markets; that is, each market is subject to a different set of supply and demand conditions.

While hundreds of house price hedonic studies have been conducted, the *identification issue* – whether we believe the parameter estimates are accurate – is an issue that continues to dog the hedonic approach. In response, a variety of estimation techniques have been used. These include ordinary least squares, two-stage least squares, other instrumental variable techniques, spatial regression approaches, the borders approach, and the mixed index approach. Traditionally, ordinary least squares has been used to estimate house price hedonics. One example is Hoehn, Berger, and Blomquist (1987), which deserves special note in this chapter. This paper uses the hedonic method not only to estimate a house price equation but also to estimate a wage equation, an additional application of the hedonic technique. Furthermore, it helps bridge the gap between the hedonic and residential mobility literatures by including a considerable number of quality of life controls including numerous measures of pollution, climate, and other urban amenities.

Subsequent research recognized the drawbacks of ordinary least squares for hedonic regressions. An early adopter of instrumentation is Voith (1991), who simultaneously runs wage and rent regressions using two-stage least squares. Instrumentation is necessary, Voith says, because wages and rents are simultaneously determined. Another attempt to identify house price hedonics using instrumental variables is found in Epple and Sieg (1999).

Although not a statistical method, another approach to identifying parameter estimates from hedonic models is to exploit borders. The technique was pioneered by Gill (1983) and was all but forgotten until the late 1990s, since which time it has become more popular than ever (although Gill's paper is rarely credited).

The idea is to examine houses that are similar in all respects but one: the houses on one side of a border have characteristic x, and those on the other side of the border have characteristic y. The border could be a school district boundary, as in Gill (1983), or it could be a state line, county line, tax abatement zone boundary, or any other boundary of interest to researchers. Gill picks areas with similar housing that have less than 10 % of blacks in neighborhood schools. Some areas are within the Columbus (Ohio) City School District; others are in suburban school districts. Suddenly in 1978, a school busing plan was approved that mandated all city schools to have between 20.9 % and 50.9 % black students. Suburban schools were exempt from busing. Gill finds evidence of increased demand for suburban houses, especially those with four or more bedrooms, when the courts first ordered city schools to desegregate.

Despite the current popularity of the *borders approach* to identifying house price hedonics, LaCombe (2004) shows that the use of spatial statistics is better. Note that while spatial statistics is not a technique in itself, it is an approach that incorporates spatial dependence in a variety of regression techniques. LaCombe takes counties that border each side of a state. He studies whether state-level differences in AFDC and food stamp usage is associated with a different prevalence of female-headed households and labor force participation levels by women. LaCombe uses both the borders approach and *spatial statistics*. While the borders approach shows no relationship between AFDC and food stamps on the percentage of female-headed households, the spatial statistical models show a positive, statistically significant relationship, a finding more consistent with theory and previous studies. LaCombe argues that the driver of this result is the presence of spatial dependence: the idea that observations nearby in space are similar to each other but that observations farther away from each other are less similar. The same Brasington and Hite (2005) paper described above for two-stage hedonic demand uses spatial econometrics to address spatial dependence. The specific procedure used in Brasington and Hite is the *spatial Durbin model*. Brasington and Hite describe the intuition behind how their spatial Durbin model incorporates the influence of *omitted variables*, thus helping to identify a house price hedonic (p. 63):

while a lagged dependent variable in time series regressions relies on observations nearby in time, the spatial lag relies on a linear combination of house values nearby in space. Unmeasured influences help determine the value of neighboring houses and, as explained earlier, the value of neighboring houses is related to the value of our own house. So our own house value is affected by the unmeasured influences of neighboring observations. And the unmeasured influences of neighboring houses are similar to the unmeasured influences for our house because our neighbors are close: the same things that affect our neighbors should affect us, too.

More formal and detailed discussions of spatial models and omitted variable bias are given in Fischer and Getis (2010) and LeSage and Pace (2010; ► [Chap. 77, “Interpreting Spatial Econometric Models”](#)). There are many models within spatial econometrics and spatial statistics that can incorporate the influence of spatial dependence. These include the spatial autoregressive model, the general

spatial model, the geographically weighted regression, and the spatial Durbin model. Certain models incorporate the influence of omitted variables; others do not. Spatial routines are available in Matlab at www.spatial-econometrics.com and www.spatial-statistics.com and also with S+, Spatial Stata, PySAL, STARS, Open GeoDa, GeoR, and GeoDa Space, most of which are free.

A final technique to identify hedonics comes from an idea of Bowden (1992), operationalized by Brasington and Hite (2008). In their introduction, Brasington and Hite (2008) discuss the shortcomings of the borders approach, instrumental variables approach, panel data approaches, and spatial statistical approaches. They then discuss Bowden's observation that Tiebout sorting leads to endogeneity of house price hedonic variables and that even including buyer characteristics as explanatory variables is insufficient to identify the hedonic. What is needed, Bowden says, is a *mixed index model*. This mixed index model consists of a system of equations estimated simultaneously. This system contains a house price hedonic with explanatory variables for house characteristics and buyer characteristics (Brasington and Hite add neighborhood characteristics to Bowden's list). The system of equations also contains equations with dependent variables representing endogenous explanatory variables from the house price hedonic. Brasington and Hite implement Bowden's mixed index model for the first time. The endogenous variables they choose from the house price hedonic are the homeowner's income, commute time, and air pollution in the homeowner's census block group. These endogenous variables are regressed as a function of additional homeowner characteristics, house characteristics, and neighborhood characteristics. Brasington and Hite find that the mixed index model yields predicted house prices that are statistically significantly different from a traditional ordinary least squares regression and from a traditional regression that adds buyer characteristics. They find the magnitude of the *capitalization of environmental quality* differs a lot when using the mixed index model, and they find the mixed index model has the most accurate policy predictions with the lowest prediction variance and the most favorable skewness and kurtosis performance. The major drawback for implementing the mixed index model is that data on the characteristics of the buyers of each house is rarely available. Furthermore, even the most thoughtful exclusion restrictions in the system of equations are somewhat arbitrary. Still, the favorable statistical properties may make the mixed index model an attractive choice when data on individual *buyer characteristics* becomes more widely available.

9.4 Conclusions

This chapter has examined the state of the literature on residential mobility and house price hedonics. It has discussed the theoretical framework of the literature, looked at the empirical approaches to test theory, and briefly discussed the factors influencing mobility and house prices.

While the theory behind the second-stage demand regressions stemming from house price hedonics continues to receive some attention, the theory behind residential mobility and first-stage house price hedonics seems relatively stable. Rather, it is the empirical side of the story that has received much more exploration in recent decades.

The residential mobility literature has advanced far in testing the determinants of mobility, but less energy has been invested in the empirical techniques used. While traditional probits and the like are appropriate and certain studies like Partridge and Rickman (2003) have applied structural vector autoregression to traditional questions, there is room to apply newer econometric approaches to investigate long-standing issues in the literature. For example, in a discipline in which space plays such a prominent role, there is a notable lack of research applying spatial statistics and econometrics to the residential mobility literature (Fischer and Getis, 2010). There is also room to investigate other determinants of mobility that have been understudied. For instance, what is the impact of having a city-county government on in-migration and out-migration? Is moving more or less common than in urban areas with a large number of highly fragmented political jurisdictions?

Concerning the house price hedonic literature, there has been some advance in regression technique, with several approaches available to help identify parameter estimates in first- and second-stage hedonic regressions. What is needed is more research like Lacombe (2004) to *compare the different techniques* to see which offer the best combination of desirable estimation qualities like unbiasedness and efficiency. The hedonic literature is a major beneficiary of renewed interest in *housing prices* as much of the developed world suffers through a post-bubble *housing bust*. Some of this renewed literature could investigate additional hedonic determinants of rises and falls in housing prices, including the role of climate and views of lakes and oceans and the role of the industries a city specializes in – the hedonic counterpart to Rappaport (2007) and Partridge and Rickman (2003).

Both the residential mobility and the house price hedonic literatures could benefit from a comprehensive data set that would put the Nurses' Health Study and the Panel Study of Income Dynamics to shame. Such a data set would need to cover an entire continent to account for mobility of subjects. It would need to enroll a large number of subjects. All the purchases of these subjects would need to be tracked, as well as the sales of durable goods. The residences and workplaces of these subjects throughout their lifetimes would be recorded, as would health, fertility, and income data. And detailed information about the subjects' residences and neighborhoods would be collected, as well as certain information about the subjects' relatives. The information about relatives would be useful for constructing instrumental variables for endogenous variables. Data on education, religious practice, criminal records, and misbehavior in school would also be useful. Data collected on such persons over the course of a lifetime would prove a research gold mine, if confidentiality could be maintained.

References

- Bowden RJ (1992) Competitive selection and market data: the mixed-index problem. *Rev Econ Stud* 59:625–633
- Brasington DM (2002) Edge versus center: finding common ground in the capitalization debate. *J Urban Econ* 52:524–541
- Brasington DM, Hite D (2005) Demand for environmental quality: a spatial hedonic analysis. *Reg Sci Urban Econ* 35:57–82
- Brasington DM, Hite D (2008) A mixed index approach to identifying hedonic price models. *Reg Sci Urban Econ* 38:271–284
- Brown JN, Rosen HS (1982) On the estimation of structural hedonic price models. *Econometrica* 50:765–768
- Brueckner JK (2011) Lectures on urban economics. MIT Press, Cambridge
- Cheshire P, Sheppard S (1998) Estimating the demand for housing, land, and neighbourhood characteristics. *Oxford Bull Econ Stat* 60:357–382
- Clark WAV, Davies-Withers S (1999) Changing jobs and changing houses: mobility outcomes of employment transitions. *J Reg Sci* 39:653–673
- Clark WAV, van Lierop WFJ (1986) Residential mobility and household location modelling. In: Nijkamp P (ed) *Handbook of regional and urban economics*, vol 1, Regional economics. North-Holland, Amsterdam, pp 97–132
- Dieleman FM (2001) Modelling residential mobility: a review of recent trends in research. *J Hsg Built Environment* 16:249–265
- Ekeland I, Heckman JJ, Nesheim L (2002) Identifying hedonic models. *Am Econ Assoc Papers Proc* 92:304–309
- Epple D, Sieg H (1999) Estimating equilibrium models of local jurisdictions. *J Polit Econ* 107:645–681
- Faggian A, McCann P, Sheppard S (2007) Some evidence that women are more mobile than men: gender differences in U.K. graduate migration behavior. *J Reg Sci* 47:517–539
- Ferreira F (2010) You can take it with you: proposition 13 tax benefits, residential mobility, and willingness to pay for housing amenities. *J Pub Econ* 94:661–673
- Fischer MM, Aufhauser E (1988) Housing choice in a regulated market. A nested multinomial logit analysis. *Geogr Anal* 20:47–69
- Fischer MM, Getis A (2010) *Handbook of applied spatial analysis. Software tools, methods and applications*. Springer, Berlin/Heidelberg/New York
- Gatzlaff DH, Ling DC (1994) Measuring changes in local house prices: an empirical investigation of alternative methodologies. *J Urban Econ* 35:221–244
- Gill HL (1983) Changes in city and suburban house prices during a period of expected school desegregation. *Southern Econ J* 50:169–184
- Hanushek E, Quigley J (1978) An explicit model of intrametropolitan mobility. *Land Econ* 54:411–429
- Haurin DR, Hendershott PF (1991) House price indexes: issues and results. *AREUEA J* 19:259–269
- Hoehn JP, Berger MC, Blomquist GC (1987) A hedonic model of interregional wages, rents, and amenity values. *J Reg Sci* 27:605–620
- Kan K (2007) Residential mobility and social capital. *J Urban Econ* 61:436–457
- Lacombe D (2004) Does econometric methodology matter? An analysis of public policy using spatial econometric techniques. *Geogr Anal* 36:105–118
- LeSage J, Fischer MM (2010) Spatial econometric methods for modeling origin-destination flows. In: Fischer MM, Getis A (eds) *Handbook of applied spatial analysis*. Springer, Berlin/Heidelberg/New York, pp 409–433
- LeSage J, Pace RK (2010) *Introduction to spatial econometrics*. CRC Press, Boca Raton
- Oates WE (1969) The effects of property taxes and local public spending on property values: an empirical study of tax capitalization and the Tiebout hypothesis. *J Polit Econ* 77:957–971

- Orr LL (1968) The incidence of differential property taxes on urban housing. *Nat Tax J* 21:253–262
- Partridge MD, Rickman DS (2003) The waxing and waning of regional economies: the chicken-egg question of jobs versus people. *J Urban Econ* 53:76–97
- Rabe B, Taylor M (2010) Residential mobility, quality of neighbourhood and life course events. *J R Stat Soc Ser A Stat Soc* 173:531–555
- Rappaport J (2007) Moving to nice weather. *Reg Sci Urban Econ* 37:375–398
- Rosen S (1974) Hedonic prices and implicit markets: product differentiation in pure competition. *J Polit Econ* 82:34–55
- Rossi PH (1955) Why families move: a study in the social psychology of urban residential mobility. Free Press, Glencoe
- Tiebout C (1956) A pure theory of local expenditures. *J Polit Econ* 64:416–424
- Voith R (1991) Capitalization of local and regional attributes into wages and rents: differences across residential, commercial, and mixed-use communities. *J Reg Sci* 31:127–145
- Winstanley A, Thorns DC, Perkins HC (2002) Moving house, creating home: exploring residential mobility. *Housing Stud* 17:813–832

Section II

Regional Economic Growth

The diagram is a word cloud centered around the concept of growth. The word 'growth' is the largest and most prominent word, rendered in a large red font. Surrounding it are numerous other words in various sizes and colors (red, orange, yellow, green, blue) representing different economic and social factors. These include 'knowledge', 'model', 'infrastructure', 'labor', 'firms', 'manufacturing', 'convergence', 'equilibrium', 'innovation', 'output', 'income', 'human spillovers', 'migration', 'investment', 'wages', 'demand', 'public', 'social', 'population', 'value', 'location', 'policy', 'consumption', 'returns', 'skills', 'well-being', 'goods', 'local', 'worker', 'firm', 'development', 'productivity', 'network', 'countries', 'cities', 'space', 'change', 'price', 'geography', 'positive', 'market', 'cost', 'cities', 'space', and 'change'. The words are arranged in a roughly circular pattern around the central 'growth' word, with larger words generally having more associated smaller words.

Maria Abreu

Contents

10.1	Introduction	170
10.2	The Solow-Swan Model	170
10.2.1	Neoclassical Production Function	171
10.2.2	Capital Accumulation Equation	173
10.2.3	Labor Supply Function	173
10.2.4	Solow Diagram and Dynamics of the Model	174
10.3	Extensions of the Solow-Swan Model	177
10.3.1	Technological Progress	177
10.3.2	Human Capital	180
10.3.3	Migration	182
10.4	The Ramsey-Cass-Koopmans Model	185
10.4.1	Households	185
10.4.2	Firms	187
10.4.3	Equilibrium and Transitional Dynamics	188
10.4.4	Extensions of the Ramsey-Cass-Koopmans Model	189
10.5	Conclusions	190
	References	190

Abstract

This chapter provides an overview of the literature on neoclassical growth, starting with the simple Solow-Swan model and highlighting the main components of the neoclassical growth process. It considers the assumptions, predictions, and limitations of the Solow-Swan model and discusses several extensions that address some of these limitations and, in particular, those that are unrealistic for a regional growth setting. Several more complex models are presented and

M. Abreu

University of Cambridge, Cambridge, UK

e-mail: ma405@cam.ac.uk

discussed, including a model that allows for exogenous technological progress, one that includes a broader definition of capital to also encompass human capital, and one that relaxes the assumption of a closed economy. Finally, the chapter considers a more complex model of neoclassical growth, the Ramsey-Cass-Koopmans model, which incorporates consumer behavior and allows for an endogenously determined savings rate.

10.1 Introduction

The publication of two seminal papers by Robert Solow, “A Contribution to the Theory of Economic Growth” in 1956 and “Technical Change and the Aggregate Production Function” in 1957, revolutionized the study of macroeconomics and started a major new research area in economic theory (Boianovsky and Hoover 2009). The neoclassical models, so-called because of their assumption of a neoclassical production function with constant returns to scale, remained the standard in growth theory until the 1980s, when new modeling techniques led to a second revolution in the field, and the advent of endogenous growth theory. However, to this day, the neoclassical models of growth retain considerable interest in the field due to their simplicity and empirical explanatory power.

This chapter provides an overview of the neoclassical models of growth, starting with the simple Solow-Swan model with exogenous savings. The model has important implications for the ultimate causes of growth and the transitional dynamics of an economy towards its long-run equilibrium or steady state. Next, the chapter discusses three important extensions of the Solow-Swan model with particular relevance to regional economies: an extension to incorporate exogenous technological progress, a version of the model that allows for physical and human capital, and an extension that relaxes the assumption of a closed economy, with no factor mobility. Finally, we will consider the Ramsey-Cass-Koopmans model, a neoclassical model that allows for endogenous savings, further relaxing the assumptions of the Solow-Swan model. The approach taken throughout the chapter is to present the models using both mathematical equations and diagrams, and the reader is directed to sources of further reading for more extensive details and discussions of other versions of the models.

The remainder of the chapter is organized as follows: Sect. 10.2 discusses the main characteristics of neoclassical growth models and presents the simple Solow-Swan model. Section 10.3 highlights several important extensions of the original model and discusses their strengths and limitations. Section 10.4 presents the more complex Ramsey-Cass-Koopmans model. Section 10.5 concludes.

10.2 The Solow-Swan Model

The Solow-Swan model is named after two influential papers on economic growth published independently by Solow (1956) and Swan (1956). It is based on three

equations: a neoclassical production function, a capital accumulation equation, and a labor supply function. We discuss the three equations and their properties in turn.

10.2.1 Neoclassical Production Function

The key to understanding neoclassical growth models is the feature that sets them apart from other streams of the growth literature: the neoclassical production function. To see why, consider the production function $Y = F(K, L)$, where Y is output, K is capital, and L is labor. The function is said to be neoclassical if a number of properties hold.

First, the function must exhibit constant returns to scale, also known as homogeneity of degree one. This property holds if when increasing all inputs by a factor λ the result is an increase in output by the same factor. For instance, a business running a factory might build a second factory and operate it using the same amount of capital and labor as the first. If the production process is based on constant returns to scale technology, this will result in a doubling of the output. Formally, this assumption can be stated as

$$F(\lambda K, \lambda L) = \lambda F(K, L) \quad (10.1)$$

Second, there must be positive and diminishing returns to all inputs. This holds if adding an additional unit of each input to the production process results in a positive increase in output, but the increase becomes smaller as more of the input is added, while holding all other inputs constant. Formally we have

$$\begin{aligned} \frac{\partial F}{\partial K} &> 0, & \frac{\partial^2 F}{\partial K^2} &< 0 \\ \frac{\partial F}{\partial L} &> 0, & \frac{\partial^2 F}{\partial L^2} &< 0 \end{aligned} \quad (10.2)$$

Finally, two sets of technical conditions must hold. The so-called Inada conditions state that the marginal product of capital will approach infinity as the labor input tends to zero and approach zero as the labor input tends to infinity (and vice versa for labor). In addition, both inputs must be *essential*, that is, a strictly positive amount of each input is needed to produce a unit of output (Barro and Sala-i-Martin 2004; Acemoglu 2009).

The above conditions define the neoclassical production function. Because the focus of growth is generally on per-capita values (such as GDP per capita), the production function is often expressed in per-capita terms. Since the neoclassical production function is subject to constant returns to scale, we can divide all the variables by the same scaling factor $\lambda = 1/L$ to give

$$Y/L = F(K/L, L/L) = F(K/L, 1) \quad (10.3)$$

We can express all the variables in their *per-capita* terms, also known as in *intensive form*, to give

$$y = f(k) \quad (10.4)$$

where y is output per capita and k is an input capital per capita. Since the Solow-Swan model is a model of the macroeconomy, the output is a composite variable, generally taken to mean total output in the economy. In practical terms, this is often operationalized using gross domestic product (GDP) in the case of countries or gross value added (GVA) in the case of regions. Writing the neoclassical production function out in intensive form also highlights one of its most important properties, which is the absence of scale effects. Production in large countries or regions is not affected by the size of the market, but rather is determined by the amount of physical capital that is available to each worker.

Another important aspect of this production function concerns the marginal products of the factors of production. These are given by the derivatives of Eq. (10.4) with respect to capital and labor:

$$\begin{aligned} \frac{\partial F}{\partial K} &= f'(k) \\ \frac{\partial F}{\partial L} &= f(k) - kf'(k) \end{aligned} \quad (10.5)$$

In a perfectly competitive economy, the factors of production are each paid their marginal product. In other words, firms will hire labor until the marginal product of labor is equal to the wage rate w and will rent capital until the marginal product of capital is equal to the rental price R .

A commonly used production function is the Cobb-Douglas function, given by

$$Y = K^\alpha L^{1-\alpha} \quad (10.6)$$

where α is a constant such that $0 < \alpha < 1$. Expressing the function in intensive form gives

$$y = k^\alpha \quad (10.7)$$

The Cobb-Douglas production function satisfies all the properties of neoclassical production functions as discussed above and also has an additional desirable property, which is that the factor income shares, or the proportions of output that accrue to each factor of production, are constant and equal α for capital and $(1 - \alpha)$ for labor. Formally,

$$\begin{aligned} R &= \frac{\partial F}{\partial K} = \alpha \frac{Y}{K} \\ w &= \frac{\partial F}{\partial L} = (1 - \alpha) \frac{Y}{L} \end{aligned} \quad (10.8)$$

It is easy to verify that the payments to the factors of production exhaust total output since

$$wL + RK = \alpha Y + (1 - \alpha)Y = Y \quad (10.9)$$

10.2.2 Capital Accumulation Equation

The second equation of the Solow-Swan model describes changes in the capital stock over time. In the discussion that follows, we will make use of *dot* notation to denote time derivatives, so that $\partial K / \partial t = \dot{K}$, and similarly for other variables. The change of the capital stock, K , over time is given by

$$\dot{K} = sF(K, L) - \delta K \quad (10.10)$$

where s is the savings rate (which is equal to the investment rate) and δ is the capital depreciation rate or the rate at which the existing capital stock becomes obsolete. The model thus assumes that all of the income that is not consumed (and is therefore *saved*) is invested in new capital stock. Equation (10.10) can also be expressed in intensive form, by using the relationship $k = K/L$, taking logs and differentiating with respect to time to give a relationship between growth rates:

$$\frac{\dot{k}}{k} = \frac{\dot{K}}{K} - \frac{\dot{L}}{L} \quad (10.11)$$

Dividing Eq. (10.10) by K and substituting into Eq. (10.11) give

$$\frac{\dot{k}}{k} = \frac{sF(K, L)}{K} - \delta - \frac{\dot{L}}{L} \quad (10.12)$$

and multiplying both sides by $k = K/L$ and rearranging give

$$\dot{k} = sf(k) - (\delta + \frac{\dot{L}}{L})k \quad (10.13)$$

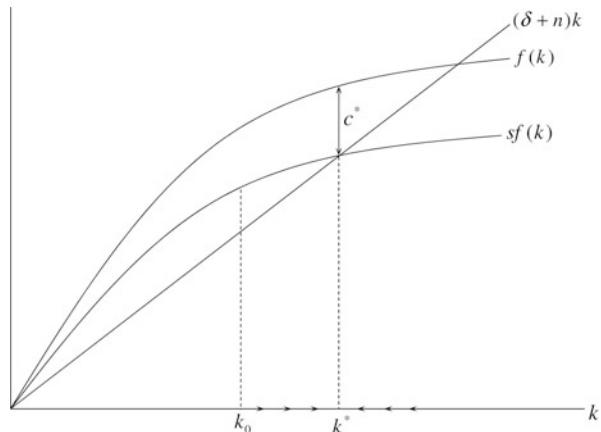
The final component of the model is the so far undefined labor force growth rate, given by \dot{L}/L , which is the final equation of the model.

10.2.3 Labor Supply Function

The Solow-Swan model in its simplest form assumes a constant labor force growth rate, so that

$$L = L_0 e^{nt} \Rightarrow \frac{\dot{L}}{L} = n \quad (10.14)$$

Fig. 10.1 The Solow Diagram



where L_0 is the initial supply of labor and n is the labor force growth rate. The model thus assumes that the labor force changes at a constant rate, given by the exogenous n . Substituting the expression for the growth rate of the labor force Eq. (10.14) into Eq. (10.13) gives the capital accumulation equation in intensive form:

$$\dot{k} = sf(k) - (\delta + n)k \quad (10.15)$$

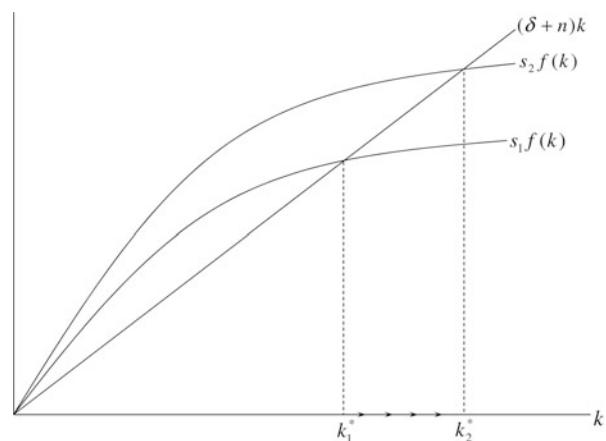
10.2.4 Solow Diagram and Dynamics of the Model

Equation (10.15) is the fundamental differential equation of the model and describes the evolution of the capital stock as a function of capital per worker k and a set of constant factors. The term $(\delta + n)$ can be thought of as the effective depreciation rate of the capital-labor ratio. In other words, the capital stock per worker declines in part because of depreciation (which makes a fraction of existing capital obsolete) and in part because of the labor force growth rate.

The workings of the model can be shown most effectively on a phase diagram, an innovative method introduced by Solow (1956), and which is now known as the *Solow diagram* (see Fig. 10.1). The diagram shows the neoclassical production-function $f(k)$ and the two components of the capital accumulation equation given by Eq. (10.15). The curve $sf(k)$ shows the amount of investment in capital per worker and has the same shape as the production function but is scaled down by a factor s . The line $(\delta + n)k$ shows the effective depreciation in capital per worker.

In order to analyze the predictions of the model, we first consider the long-run or *steady state* behavior and then discuss the short-run dynamics. In the long run, the model predicts that an economy will reach a steady state where all the variables are either constant or growing at a constant, steady rate. The steady state occurs when the capital stock per worker is constant, so that $\dot{k} = 0$. As can be seen from Eq. (10.15), this is only possible when the level of investment per worker just

Fig. 10.2 The Impact of a Higher Savings Rate



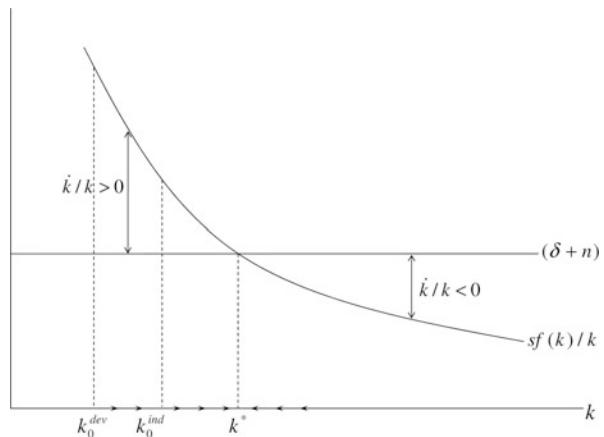
equals the effective depreciation of capital per worker, given by the point where the curves $sf(k)$ and $(\delta + n)k$ meet. At this point, the steady state capital stock is equal to k^* in Fig. 10.1. To see why this is a steady state, consider an economy that initially has a lower capital stock per worker, such as k_0 . At this point, the investment in new capital stock per worker $sf(k)$ is greater than the effective depreciation per worker $(\delta + n)k$ so that the capital stock per worker is growing and will continue to grow until the steady state is reached. Similarly, if the economy is at a point to the right of k^* , effective depreciation exceeds new investment, and the capital stock per worker falls until the steady state is reached. Note that in the steady state, the capital stock per worker is constant, while the level of the capital stock K grows at the constant rate of labor growth n . Similarly, output per worker $y^* = f(k^*)$ and consumption per worker $c^* = (1 - s)f(k^*)$ are constant, while output Y and consumption C grow at the constant rate n .

The Solow diagram can also be used to study the effects of a change in one of the fundamental factors of the model, for instance, the savings rate s . As shown on Fig. 10.2, an increase in the savings rate from s_1 to s_2 results in an upward shift of the $sf(k)$ curve and, as a result, a higher steady state level of capital per worker k_2^* . However, the effect of this increase on growth is only temporary. The economy initially grows as the level of capital per worker increases, but as the new steady state is reached, capital and output per worker cease to grow and become constant, while the new levels of capital and output grow steadily at the rate of the labor force growth n . To summarize, an economy with a higher rate of saving will achieve a higher level of output per worker but will, once the steady state is reached, grow at the same rate as an economy with a lower savings rate.

The equation for the growth rate of capital per worker can be seen by dividing the capital accumulation equation given by Eq. (10.15) by k :

$$\frac{\dot{k}}{k} = \frac{sf(k)}{k} - (\delta + n) \quad (10.16)$$

Fig. 10.3 Neoclassical Convergence



The implications of the dynamics of the growth rate of capital can also be seen diagrammatically by drawing two curves corresponding to the two components of Eq. (10.16), shown on Fig. 10.3. The growth rate of capital per worker is zero in the steady state, when the two curves intersect and capital per worker is equal to k^* . At other values of k , the growth rate is given by the gap between the two curves, so the furthest an economy is from its steady state, the fastest it grows toward it. In other words, an economy grows asymptotically toward its steady state. The reason for this result is the assumption of diminishing returns to capital of the neoclassical production function; when k is relatively low, the average product of capital is high, and hence, the gross investment per unit of capital $sf(k)/k$ is high, while the effective rate of depreciation stays constant. The growth rate is positive if the level of capital per worker is lower than the steady state and negative if the current level is higher than the steady state.

This last result has important implications for the comparative study of growth rates across countries or regions. Specifically, as long as two economies have the same production function and the same savings, labor force growth, and capital depreciation rates, they will converge to the same steady state. Differences in growth rates will only be observed if the two economies are at different stages in this convergence process, for instance, because one economy is less developed than the other and has a smaller initial value of k . As shown on Fig. 10.3, the less developing economy has a lower level of capital per worker, denoted k_0^{dev} , while the more industrialized economy initially has k_0^{ind} . While both economies are converging to the same steady state k^* , the developing economy will temporarily have a faster rate of growth, while it *catches up* to the capital per-worker level of the industrialized economy.

The speed of convergence of a country to its steady state is thus a measure of how fast the growth rate declines as the capital stock increases or formally

$$\beta = - \frac{\partial(\dot{k}/k)}{\partial \log k} \quad (10.17)$$

where β is the speed of convergence. A large empirical literature exists to estimate the speed of convergence, typically in a cross section of countries or regions. Note that, as discussed above, the concept of convergence only applies to economies converging to their own steady states, unless all countries or regions that are included in the analysis have the same production function, savings, population growth, and capital depreciation rates. This is more likely to be the case for regions within the same country or countries that are relatively similar (e.g., the OECD economies), although population growth rates and the level of technology can sometimes vary markedly across regions within one country. Many empirical studies control for differences in these factors by including them in a multivariate regression context.

The simple Solow-Swan model discussed up until now abstracts from reality in several important ways. First, it does not allow for technological progress and assumes that the same combinations of capital and labor will always result in the same amount of output. Second, it assumes that all units of labor are equally productive, regardless of their level of skill. The model thus abstracts from education and training. Finally, and importantly, in the context of regions, the model assumes a closed economy, that is, there are no exports, and capital and labor are assumed immobile. We next discuss several extensions of the Solow-Swan model to incorporate these factors. Note that we have so far abstracted from markets and the behavior of individual households and firms. We will discuss these micro-foundations of the model in more detail in Sect. 10.4. However, it is worth mentioning at this point that the results of the simple Solow-Swan model can also be shown to hold in a framework that explicitly incorporates markets for labor and financial assets (see Barro and Sala-i-Martin 2004, pp. 31–33 for details).

10.3 Extensions of the Solow-Swan Model

10.3.1 Technological Progress

Introducing exogenous technological progress into the Solow-Swan model is fairly straightforward. The first step is to include the level of technology in the neoclassical production function. This can take several forms, depending on whether we define the technology as a process that allows production of the same amount of output with less capital input (capital-saving technology) and less labor input (labor-saving technology) or a process that does not save relatively more of either input (neutral or unbiased technological progress). The definition of the latter also varies. For the purposes of the Solow-Swan model, we assume a *Harrod neutral* or *labor-augmenting* technology, which is necessary to ensure that the model has a steady state (see Barro and Sala-i-Martin 2004, pp. 52–53 for a discussion). Our production function now takes the form

$$Y = F(K, AL) \tag{10.18}$$

where A is the level of labor-augmenting technology, which raises output in the same way as an increase in the stock of labor. We therefore refer to AL as the *effective units of labor* used in production. A distinguishing characteristic of the neoclassical models is that technological progress is assumed to be exogenous, so that

$$A = A_0 e^{gt} \Rightarrow \frac{\dot{A}}{A} = g \quad (10.19)$$

where g is the exogenous growth rate of technology and A_0 is the initial level of technology. The capital accumulation in Eq. (10.10) still holds, but we construct a new *state* variable $\tilde{k} = K/AL$, defined as *capital per effective unit of labor*, and which is constant in the steady state. Taking logs and differentiating with respect to time, we obtain

$$\frac{\dot{\tilde{k}}}{\tilde{k}} = \frac{\dot{K}}{K} - \frac{\dot{A}}{A} - \frac{\dot{L}}{L} \quad (10.20)$$

Substituting Eqs. (10.10), (10.14), and (10.19) into Eq. (10.20) gives

$$\frac{\dot{\tilde{k}}}{\tilde{k}} = \frac{sF(K, AL)}{K} - \delta - g - n \quad (10.21)$$

and rearranging Eq. (10.21) gives

$$\dot{\tilde{k}} = \frac{sF(K, AL)}{AL} - (\delta + g + n)\tilde{k} \quad (10.22)$$

The production function can be also be expressed in terms of effective units of labor, so that $\tilde{y} = f(\tilde{k})$. Substituting into Eq. (10.22) gives the capital accumulation equation for the Solow-Swan model, in terms of effective units of labor:

$$\dot{\tilde{k}} = sf(\tilde{k}) - (\delta + g + n)\tilde{k} \quad (10.23)$$

To derive the steady state of the model, we now proceed as before. The steady state occurs when the capital per effective unit of labor is constant or $\dot{\tilde{k}} = 0$. This occurs when

$$sf(\tilde{k}) = (\delta + g + n)\tilde{k} \quad (10.24)$$

that is, when the gross investment per unit of effective labor is just equal to the reduction in capital per unit of effective labor due to the effective depreciation rate that now includes technological progress. In the long run, therefore, capital and output *per unit of effective labor* are constant, capital and output *per worker* grow at the constant rate of technological progress g , and capital and output grow at the constant rate $(n + g)$. As with the simple model discussed above, the short-run

growth rate in capital per worker (and therefore in output per worker) is greater the further a country or region is from its steady state. During transition, the growth rate in output per worker is therefore greater than the rate of technological progress, declining asymptotically until reaching the constant growth rate g in the steady state.

The speed of convergence of an economy toward its own steady state was defined in Eq. (10.17) as the proportional decrease in the growth rate of the state variable as the capital stock declines. To see how it varies with the variables of the augmented model, we first assume that the economy has a Cobb-Douglas production function with labor-augmenting technology given by

$$Y = F(K, AL) = K^\alpha(AL)^{1-\alpha} \quad (10.25)$$

or $y = \tilde{k}^\alpha$ in augmented form, where α is a constant such that $0 < \alpha < 1$. By dividing Eq. (10.23) by \tilde{k} , we obtain the growth rate of capital per unit of effective labor:

$$\frac{\dot{\tilde{k}}}{\tilde{k}} = \frac{sf(\tilde{k})}{\tilde{k}} - (\delta + g + n) \quad (10.26)$$

In the Cobb-Douglas case, this is given by

$$\frac{\dot{\tilde{k}}}{\tilde{k}} = \frac{s\tilde{k}^\alpha}{\tilde{k}} - (\delta + g + n) = s\tilde{k}^{-(1-\alpha)} - (\delta + g + n) \quad (10.27)$$

We can rewrite the expression for the growth rate of \tilde{k} given in Eq. (10.27) as a function of $\log(\tilde{k})$

$$\frac{\dot{\tilde{k}}}{\tilde{k}} = se^{-(1-\alpha)\log(\tilde{k})} - (\delta + g + n) \quad (10.28)$$

and differentiate with respect to $\log(\tilde{k})$ to give an expression for the speed of convergence β , as defined in Eq. (10.17)

$$\beta = -\frac{\partial(\dot{\tilde{k}}/\tilde{k})}{\partial \log \tilde{k}} = (1 - \alpha)s\tilde{k}^{-(1-\alpha)} \quad (10.29)$$

At the steady state, $s\tilde{k}^{-(1-\alpha)} = (\delta + g + n)$, as shown by Eq. (10.27). Therefore, in the neighborhood of the steady state, the speed of convergence is given by

$$\beta^* = (1 - \alpha)(\delta + g + n) \quad (10.30)$$

The model therefore predicts that the rate of long-run per-capita output growth is equal across all countries or regions, since it is given by the exogenous rate of technological progress. Because technological progress is determined outside the

model (or, in other words, is not explained by the model), the model is silent on the ultimate causes of long-run growth. This shortcoming, in particular, has resulted in a large new stream of literature, known as *endogenous growth* or *new growth*, that has sought to explain the origins of technological progress.

10.3.2 Human Capital

A second major limitation of the simple Solow-Swan model is its abstraction from education and skills, although these variables clearly differ across countries (and regions). It is therefore no surprise that an extension of the Solow-Swan model to incorporate human capital came about for empirical reasons, as it became apparent that the original model failed to accurately explain differences in growth rates and, in particular, to adequately explain the speed of convergence given in Eq. (10.30).

For instance, Barro and Sala-i-Martin (2004) use benchmark values for the United States economy of $\delta = 0.05$ per year, $g = 0.02$ per year, and $n = 0.01$ per year. The value for the growth rate of technology corresponds to the long-run growth rate of GDP in the United States, while the values for the capital depreciation and labor force growth rate are long-run averages. They also assume a value of the capital share of income of $\alpha = 1/3$, in line with previous studies. Using these values, the predicted speed of convergence is $\beta^* = 0.053$ or 5.3 % per year, which is much larger than the observed average speed of convergence of around 2 % per year. Such a low value would be consistent with a higher capital share of income of the order of $\alpha = 0.75$.

One way to resolve this discrepancy is to allow for both human and physical capital in the model, thus leading to a higher capital share in income. We use an augmented version of the Cobb-Douglas production function given in Eq. (10.25):

$$Y = K^\alpha H^\eta (AL)^{1-\alpha-\eta} \quad (10.31)$$

where H is human capital, L is raw labor, and α and η are constants such that $0 < \alpha < 1$ and $0 < \eta < 1$. We again express the model in terms of units of effective labor by dividing all variables by AL . Output per unit of effective labor is now given by

$$\tilde{y} = f(\tilde{k}, \tilde{h}) = \tilde{k}^\alpha \tilde{h}^\eta \quad (10.32)$$

We assume, as before, that households consume a constant share of their income, with the remainder saved and invested in capital. We further make the simplifying assumption that both physical and human capital depreciate at the same rate of δ . The capital accumulation equation is then given by

$$\dot{\tilde{k}} + \dot{\tilde{h}} = sf(\tilde{k}, \tilde{h}) - (\delta + g + n)(\tilde{k} + \tilde{h}) \quad (10.33)$$

and substituting expression (10.32) into (10.33) gives

$$\dot{\tilde{k}} + \dot{\tilde{h}} = s\tilde{k}^\alpha \tilde{h}^\eta - (\delta + g + n)(\tilde{k} + \tilde{h}) \quad (10.34)$$

Households will invest in the form of capital that gives the highest rate of return, if investment in both forms of capital is positive. In equilibrium, the marginal products of physical and human capital must therefore be equal:

$$\alpha \frac{\tilde{y}}{\tilde{k}} - \delta = \eta \frac{\tilde{y}}{\tilde{h}} - \delta \quad (10.35)$$

Rearranging, we obtain a relationship between physical and human capital:

$$\tilde{h} = \frac{\eta}{\alpha} \tilde{k} \quad (10.36)$$

and substituting Eq. (10.35) into the capital accumulation equation given by Eq. (10.33) results in

$$\dot{\tilde{k}} = s\phi k^{\alpha+\eta} - (\delta + g + n)\tilde{k} \quad (10.37)$$

where $\phi = (\eta^\eta \alpha^{1-\eta})/(\alpha + \eta)$ is a constant. Following the same procedure used in Eqs. (10.26)–(10.30), we arrive at a new expression for the speed of convergence in the neighborhood of the steady state:

$$\beta^* = (1 - \alpha - \eta)(\delta + g + n) \quad (10.38)$$

Given a reasonable estimate of the human capital share of $\eta = 0.4$ and the other estimates discussed above, we obtain an estimated speed of convergence of $\beta^* = 0.21$ or 2.1 % per year, a result that is in line with much of the empirical literature.

The model presented above is a slight variation of the approach followed by Mankiw et al. (1992) in a seminal paper. The authors estimate a model with two capital accumulation equations, one for physical and one for human capital, with the rates of investment in the two types of capital being determined independently of each other. The expression for the speed of convergence obtained by Mankiw et al. (1992) is the same as that given by Eq. (10.38), and the authors find strong support for the predictions of the neoclassical model. In particular, their estimated speed of convergence is close to 2 %, particularly if the sample is restricted to similar countries such as the OECD economies (see Abreu et al. 2005, for an extended discussion). It is also worth noting that recent approaches to modeling regional growth build on the Solow model, with or without human capital, and extend it in new ways, such as by incorporating technology spillovers across countries (Ertur and Koch 2007) or regions (Fischer 2011).

10.3.3 Migration

The assumption of a closed economy, implying that the factors of production are immobile across economies, is particularly unrealistic if the model is applied to regions (rather than countries). We are therefore interested in extensions of the neoclassical approach that relax the assumptions of labor and capital mobility. In this section, we consider a version of the Solow-Swan model that allows for migration and is based on a model developed in Barro and Sala-i-Martin (2004).

We allow for the movement of labor but not capital or goods, that is, we assume there is no trade or capital mobility that is independent of migration. We do, however, assume that migrants carry capital as they move, most of which is in the form of human capital. In what follows, we do not distinguish between physical and human capital, but denote by κ the total quantity of broad capital carried by each migrant. The net flow of migrants is given by M , which is equal to in-migration minus out-migration, and the net migration rate is given by M/L . The overall growth in the domestic labor force is therefore given by

$$\frac{\dot{L}}{L} = n + m \quad (10.39)$$

Since migrants bring κM units of capital into the domestic economy, the capital accumulation equation is now given by

$$\dot{K} = sF(K, AL) - \delta K + \kappa M \quad (10.40)$$

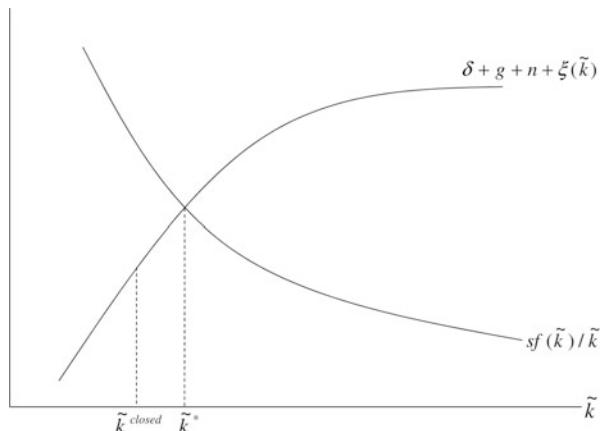
where we have assumed that technological progress is labor augmenting, as in Eq. (10.18). As before, we use Eq. (10.40) and the relationship between growth rates given in Eq. (10.20) to obtain the growth rate of capital per effective worker \tilde{k} :

$$\frac{\dot{\tilde{k}}}{\tilde{k}} = \frac{sf(\tilde{k})}{\tilde{k}} - (\delta + g + n) - m \left(1 - \frac{\tilde{\kappa}}{\tilde{k}}\right) \quad (10.41)$$

where $\tilde{\kappa} = \kappa e^{-gt}$ is the capital per effective immigrant. In the Solow-Swan model with technological progress, discussed in Sect. 10.3.1, capital per unit of effective labor depreciates due to the growth of effective labor $g + n$ and due to capital depreciation rate δ . Once we add migration to the model, there is an additional element; capital per unit of effective labor also depreciates due to migration, because migrants increase the domestic labor supply and typically have lower levels of capital than domestic workers, that is, $\tilde{\kappa} < \tilde{k}$.

We next make several assumptions about the factors included in the final term of Eq. (10.41). First, we assume that the capital per effective immigrant $\tilde{\kappa}$ is approximately constant over time, since either (a) there is in-migration, and if the typical origin country is close to its steady state, then $\tilde{\kappa} < \tilde{k}$ is independent of \tilde{k} , or (b) there is out-migration, so $\tilde{\kappa}$ is the capital per effective worker for each emigrant and $\tilde{\kappa} < \tilde{k}$ is likely to be fairly constant.

Fig. 10.4 The Solow-Swan Model with Migration



Second, we assume that there is a positive relationship between the net migration rate m and the capital per unit of effective labor \tilde{k} . This is because a higher \tilde{k} implies a higher domestic wage rate and therefore a more attractive prospect for potential migrants. The relationship between m and \tilde{k} is affected by exogenous factors such as the wage rate or standard of living in the origin countries or regions, the costs of migration, and the volume of migration.

We are now ready to analyze the conditions of the steady state and the dynamics of the model. We define the last term of the capital accumulation in Eq. (10.41) as

$$\xi(\tilde{k}) = m(\tilde{k}) \left(1 - \frac{\tilde{k}}{\tilde{k}} \right) \quad (10.42)$$

where m is now a function of \tilde{k} . Substituting Eq. (10.42) into the expression for the growth rate of capital per unit of effective worker Eq. (10.41) gives

$$\frac{\dot{\tilde{k}}}{\tilde{k}} = \frac{sf(\tilde{k})}{\tilde{k}} - [\delta + g + n + \xi(\tilde{k})] \quad (10.43)$$

It can be shown (see Barro and Sala-i-Martin 2004, p. 387, for details) that the derivative of $\xi(\tilde{k})$ with respect to \tilde{k} is positive whether the net migration rate is positive or negative. It follows that an increase in \tilde{k} raises the effective capital depreciation rate $\delta + g + n + \xi(\tilde{k})$, in contrast with previous versions of the Solow-Swan model discussed above, where this term was independent of \tilde{k} .

The implications of this can be seen on a diagram of the growth dynamics of \tilde{k} . Figure 10.4 shows the growth dynamics of the Solow-Swan model with migration. In contrast with the closed-economy case discussed in Sect. 10.2, the effective depreciation curve is now an upward sloping line, rather than a horizontal line as in Fig. 10.3. The steady state is given by \tilde{k}^* , where the two curves intersect and, as

before, the growth rate for any value of \tilde{k} is given by the vertical distance between two points. Of special interest is \tilde{k}^{closed} , which is the capital per unit of effective labor for the case where $m = 0$, and there is no migration. As drawn, the steady state level of capital per unit of effective labor \tilde{k}^* is higher than the no-migration level \tilde{k}^{closed} , so the economy will be a perpetual recipient of migrants. The diagram can be used to study the dynamics of the model. For instance, an increase in the domestic savings rate will lead to an upward shift in the $sf(\tilde{k})/\tilde{k}$ curve and therefore in higher values of \tilde{k}^* and m^* . This occurs because an increase in the savings rate raises domestic wages and makes the economy more attractive for migrants. Conversely, an exogenous increase in migration, such as that resulting from a more liberal immigration policy, leads to a lower \tilde{k}^* . This occurs because migrants arrive with relatively little capital, thus lowering the capital per effective worker in the economy.

We are also interested in the effects on the speed of convergence once migration is introduced to the model. It can be shown (see Barro and Sala-i-Martin 2004, p. 389) that the new speed of convergence is given by

$$\beta^* = (1 - \alpha - \eta)(\delta + g + n) + b + b(1 - \alpha) \log \frac{\tilde{k}^*}{\tilde{k}_{world}} \quad (10.44)$$

where the last two terms in Eq. (10.44) now incorporate the element e , defined as the derivative of $\xi(\tilde{k})$ with respect to $\log \tilde{k}$, and \tilde{k}_{world} is the capital per unit of effective labor in other regions or countries. The model therefore predicts that the speed of convergence in the model with migration is higher than the speed of convergence in the closed-economy model by the amount e , if we assume that $\tilde{k}^* = \tilde{k}_{world}$ holds for the typical economy. Empirical estimates presented by Barro and Sala-i-Martin (2004) suggest that this difference is of the order of $e = 0.003$ or 0.3 % per year for both countries and regions.

If migration is indeed an important determinant of convergence, including a measure of migration in a growth regression will lead to a lower estimate of the speed of convergence, all other things being equal. Barro and Sala-i-Martin (2004) show that the estimated speed of convergence is indeed lower once migration is accounted for, as long as the method used controls for endogeneity in the migration variable. However, the model assumes that labor is relatively homogenous, while in reality, migrants tend to be younger, better educated, and more entrepreneurial than nonmigrants (McCann 2001). An influx of migrants, especially in the regional context, can therefore lead to productivity gains in the form of technological progress, exacerbating the differences between regions and leading to divergence. In addition, since migratory workers have relatively little physical capital endowments, migration will tend to increase the rate of return to capital, leading to capital inflows. This is likely to further exacerbate the differences between receiving and sending economies. The overall effect of migration on growth may therefore be positive, negative, or insignificant, depending on the characteristics of the study (see Ozgen et al. 2010, for meta-analytical evidence).

10.4 The Ramsey-Cass-Koopmans Model

The models discussed above all have one thing in common: they are macroeconomic models that abstract from the utility maximization decisions of individual households. Thus, the choice between consumption and saving is determined exogenously. We now turn to a specification of the neoclassical model that explicitly incorporates consumer behavior with the Solow-Swan model as a special case. Our model in this section is based on Ramsey (1928), which was later refined by Cass (1965) and Koopmans (1965). For simplicity, we assume a simple production function without technological progress.

10.4.1 Households

The model assumes that households provide labor services to firms in exchange for wages, receive interest income from assets, spend a fraction of their income on the consumption of goods produced by firms, and save the remainder by accumulating assets. Households are assumed to care about their descendants, so that consumption by future members of the household enters the present utility function. Since households are assumed to be identical, we can simplify the analysis by assuming there is only one, infinitely lived, household, and that at time $t = 0$, there is only one worker in the economy. Population at time t is then given by

$$L_t = e^{nt} \quad (10.45)$$

where n is the population growth rate. Since all the income that is saved is used to accumulate assets, the household budget constraint at time t is given by

$$\dot{B} = wL + rB - C \quad (10.46)$$

where B are assets, w the wage rate, r the interest rate, and C consumption (and we have suppressed time subscripts). As before, it will be useful to express Eq. (10.46) in per-worker terms. Defining b as assets per worker, we use the relationship $b = B/L$, take logs, and differentiate to give

$$\frac{\dot{b}}{b} = \frac{\dot{B}}{B} - \frac{L}{L} \quad (10.47)$$

Substituting Eq. (10.45) and Eq. (10.46) into Eq. (10.47) and rearranging, we arrive at an expression for the change in assets per worker:

$$\dot{b} = w + rb - nb - c \quad (10.48)$$

where c is consumption per worker. At this stage, we also make the assumption that households cannot borrow unlimited amounts to finance arbitrarily high current

levels of consumption. The present value of current and future assets must therefore be positive, households cannot borrow indefinitely until the end of their economic life cycle, and household debt cannot increase at a rate asymptotically higher than the interest rate.

The intertemporal utility function of the representative household is then given by

$$\begin{aligned} U &= \int_0^\infty u(c_t) L_t e^{-\rho t} dt \\ &= \int_0^\infty u(c_t) e^{nt} e^{-\rho t} dt \end{aligned} \tag{10.49}$$

that is, the utility of the household at $t = 0$ is a weighted sum of all future flows of utility to members of the household, where $\rho > 0$ is the rate of time preference. A higher ρ implies that consumption by future members of the household is less desirable compared to current consumption. The utility function $u(c)$ gives the utility per worker and is assumed to be increasing in c , to be concave, and to satisfy the Inada conditions.

Households maximize their intertemporal utility, given by Eq. (10.49), subject to their budget constraint Eq. (10.40). Using the Hamiltonian approach, the problem is given by

$$J = u(c)e^{-(\rho-n)t} + v[w + (r - n)b - c] \tag{10.50}$$

where v is the present value of the shadow price of income in terms of utility units. The first-order conditions are

$$\begin{aligned} \frac{\partial J}{\partial c} &= 0 \Rightarrow v = u'(c)e^{-(\rho-n)t} \\ \frac{\partial J}{\partial b} &= -\dot{v} \Rightarrow \dot{v} = -(r - n)v \end{aligned} \tag{10.51}$$

and transversality condition

$$\lim_{t \rightarrow \infty} (v_t b_t) = 0 \tag{10.52}$$

The latter expression indicates that the value of the assets of the representative household must approach zero as time approaches infinity or, in other words, that households do not hold valuable assets in perpetuity. From the first-order conditions, we obtain the first fundamental equation of the model, also known as the Euler equation:

$$r = \rho - \left[\frac{u''(c)c}{u'(c)} \right] \frac{\dot{c}}{c} \tag{10.53}$$

which shows that households choose consumption so as to equate the return to savings r to the *rate of return* to consumption, the latter of which is given by the right-hand side of Eq. (10.53). In other words, households are indifferent between consumption and saving if the rates of return to the two activities are equal.

In order to parameterize the model, a commonly used functional form for the utility function is the *constant intertemporal elasticity of substitution* (CIES) function:

$$u(c) = \frac{c^{1-\theta} - 1}{1 - \theta} \quad (10.54)$$

where $\theta > 0$. The elasticity of substitution is the constant $\sigma = 1/\theta$, which reflects the willingness of individuals to accept deviations from a uniform pattern of consumption over time. Using the functional form Eq. (10.54) in the Euler equation Eq. (10.53) gives

$$\frac{\dot{c}}{c} = \frac{1}{\theta}(r - \rho) \quad (10.55)$$

Intuitively, the growth rate of consumption per worker fluctuates with the optimizing behavior of households. Households save more if the rate of return r is high relative to the rate of time preference ρ . This effect is magnified when the rate of intertemporal substitution is high, indicating that future consumption is considered a good substitute for current consumption.

10.4.2 Firms

The second side to the problem concerns the behavior of a large number of identical firms, producing a homogenous good Y . Each firm uses a neoclassical production function $Y = F(K, L)$ and pays wages w and rent R in exchange for a unit labor and capital, respectively. The net rate of return to capital is given by $R - \delta = r$, where δ is the capital depreciation rate and r is the interest rate on loans to other households. The representative firm maximizes profits given by

$$\pi = F(K, L) - (r + \delta)K - wL \quad (10.56)$$

or in terms of per-worker units

$$\pi = L[f(k) - (r + \delta)k - w] \quad (10.57)$$

The first-order conditions of the maximization problem of the firm are thus

$$\begin{aligned} \frac{\partial \pi}{\partial K} &= 0 \Rightarrow f'(k) = r + \delta \\ \frac{\partial \pi}{\partial L} &= 0 \Rightarrow f(k) - kf'(k) = w \end{aligned} \quad (10.58)$$

10.4.3 Equilibrium and Transitional Dynamics

To find the equilibrium, we combine the first-order conditions for the firm given in Eq. (10.58) with the expression for the growth rate of consumption in Eq. (10.55) to find expressions for the change of capital and consumption over time. Since in equilibrium the stock of assets per worker equals the stock of capital per worker, we have $b = k$. Starting from Eq. (10.48) and substituting $k = b$, and the expressions for w and r from Eq. (10.58), we obtain

$$\dot{k} = f(k) - (n + \delta)k - c \quad (10.59)$$

Similarly, substituting the expression for r from Eq. (10.58) into Eq. (10.55) gives

$$\frac{\dot{c}}{c} = \frac{1}{\theta}(f'(k) - \delta - \rho) \quad (10.60)$$

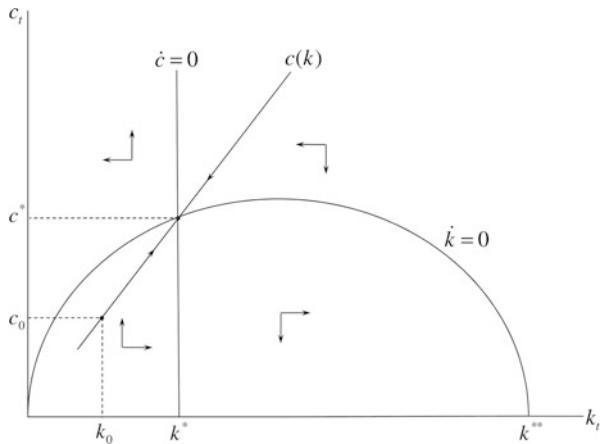
In the steady state, the growth rates of capital and consumption per worker are zero and so is the growth rate of output per worker. As with the Solow-Swan model, the simple Ramsey-Cass-Koopmans model therefore predicts that the growth of output per capita is zero in the long run. The two equations, Eqs. (10.59) and (10.60), together with the initial level of capital per worker k_0 and the transversality condition determine the time paths of capital and consumption per worker.

To see the equilibrium points of the model, we once again make use of a phase diagram. The steady state values of c and k are determined by setting expressions (10.59) and (10.60) to zero. The curve labeled $\dot{k} = 0$ on Fig. 10.5 shows the (k, c) pairs for which expression (10.59) is equal to zero, which is given by $c = f(k) - (n + \delta)k$. The peak of the curve occurs where $f'(k) - \delta = n$ or where the interest rate equals the steady state growth rate of output n , the capital per worker for which consumption is maximized. The vertical line labeled $\dot{c} = 0$ shows the (k, c) pairs for which Eq. (10.60) is set to zero. Note that the value of k for which this holds is independent of c . Since the $\dot{k} = 0$ and $\dot{c} = 0$ lines cross three times, there are three steady states. The first is at the origin, where $c = k = 0$. The second is at the point corresponding to c^* and k^* , and the third occurs at the point k^{**} where there is positive capital per worker but zero consumption.

We focus on the second steady state. The arrows on Fig. 10.5 show the direction the economy will move in for each combination of c and k . The economy can converge to the steady state given by (c^*, k^*) only if it starts in two of the four quadrants that divide the space. The dynamic equilibrium follows a saddle path shown by the diagonal line on the diagram. If the initial capital per-worker level is given by k_0 , the economy will converge to the steady state only if initial consumption per worker is equal to c_0 .

Unlike the Solow-Swan model, the Ramsey-Cass-Koopmans model predicts that the savings rate will vary with the level of development, although the exact path that it takes is complex and depends on the relative strength of income and substitution

Fig. 10.5 The Ramsey-Cass-Koopmans Model



effects. Intuitively, as k rises, a decline in $f'(k)$ lowers the rate of return on saving given by r . This reduces the incentive to save as the economy develops and constitutes an intertemporal substitution effect. However, households have a preference for consumption smoothing and therefore wish to consume a larger proportion of their income when their income is low. As k rises and they become richer, the gap between current and permanent income falls, and the savings rate tends to rise. This is the income effect. The overall behavior of the savings rate as an economy develops is therefore an empirical question. Barro and Sala-i-Martin (2004) review the empirical evidence and find that, in general, the savings rate tends to rise with per-capita income during the transition to the steady state. This in turn implies that the predicted speed of convergence is lower in the Ramsey-Cass-Koopmans model, although the growth rate in income per capita is still higher for countries or regions that are furthest away from their steady states.

10.4.4 Extensions of the Ramsey-Cass-Koopmans Model

The Ramsey-Cass-Koopmans model can be extended to allow for an open economy, a much more realistic assumption in the context of regions. This involves incorporating the mobility of goods across borders and international borrowing and lending (Barro and Sala-i-Martin 2004). However, extending the model in this way leads to extreme (and paradoxical) results. The speed of convergence becomes infinite for all but the most patient country (defined in terms of the rate of time preference), so that consumption tends to zero and assets become negative. The most patient country owns all the assets and consumes almost all output. Additional constraints, such as imperfect international credit markets, are needed to ensure less extreme predictions.

In addition, the model can be extended to allow for migration. The key relationship is now between the migration rate and the behavior of the savings rate. Migrants enter the economy with a quantity of capital, mainly in the form of

human capital. However, unlike in the standard Ramsey-Cass-Koopmans model, consumption of migrant workers does not enter into the utility function of existing residents, or in other words, existing residents do not *care* about the consumption level of migrants. If the domestic economy is attractive to migrants in its closed state, the opening up of the economy leads to a steady state with positive migration and reduced capital intensity. As with the Solow model, migration in the Ramsey-Cass-Koopmans model results in a higher speed of convergence to the steady state.

The models become increasingly complex if additional forms of spatial interaction, such as trade and technological spillovers, are also considered. In a comprehensive survey, Nijkamp and Poot (1998) discuss a range of open economy growth models that are particularly applicable to regions. For instance, introducing trade into the neoclassical models will speed up the rate of convergence and result in a pattern of specialization that reflects equilibrium factor intensities. However, if the model is extended to allow for technological spillovers, a steady state equilibrium is unlikely to exist (Nijkamp and Poot 1998). Models with technological spillovers, although no longer strictly neoclassical, can also be considered extensions of the models discussed in this chapter (Fischer 2011).

10.5 Conclusions

The chapter has provided an overview of the fundamental models of neoclassical growth and has discussed their key features, predictions, and limitations. The review starts with the simple Solow-Swan model, with an exogenous savings rate and in a closed-economy setting. This model, although simple, has considerable predictive power and provides the basis for several extensions. Several of these extensions are reviewed, among them a model that allows for exogenous technological progress, one that includes a broader definition of capital to also encompass human capital, and one that relaxes the assumption of a closed economy. Finally, the chapter considers a more complex model of neoclassical growth, the Ramsey-Cass-Koopmans model, which incorporates consumer behavior and allows for an endogenously determined savings rate. Additional extensions incorporating trade and technology diffusion are also briefly discussed.

References

- Abreu M, De Groot HLF, Florax RJGM (2005) A meta-analysis of β -convergence: the legendary 2%. *J Econ Surv* 19(3):389–420
- Acemoglu D (2009) Introduction to modern economic growth. Princeton University Press, Princeton
- Barro RJ, Sala-i-Martin X (2004) Economic growth theory. McGraw-Hill, Boston
- Boianovsky M, Hoover KD (2009) The neoclassical growth model and 20th century economics. *History Polit Econ* 41 (Supplement):1–23
- Cass D (1965) Optimum growth in an aggregative model of capital accumulation. *Rev Econ Stud* 32(3):233–240

- Ertur C, Koch W (2007) Growth, technological interdependence and spatial externalities: theory and evidence. *J Appl Econom* 22(6):1033–1062
- Fischer MM (2011) A spatial Mankiw-Romer-Weil model: theory and evidence. *Ann Reg Sci* 47(2):419–436
- Koopmans T (1965) On the concept of optimal economic growth. In: The econometric approach to development planning. North Holland, Amsterdam
- Mankiw NG, Romer D, Weil DV (1992) A contribution to the empirics of economic growth. *Quart J Econ* 107(2):407–437
- McCann P (2001) Urban and regional economics. Oxford University Press, Oxford
- Nijkamp P, Poot J (1998) Spatial perspectives on new theories of economic growth. *Ann Reg Sci* 32(1):7–37
- Ozgen C, Nijkamp P, Poot J (2010) The effect of migration on income growth and convergence: meta-analytic evidence. *Papers Reg Sci* 89(3):537–561
- Ramsey F (1928) A mathematical theory of saving. *Econ J* 38(152):543–559
- Solow RM (1956) A contribution to the theory of economic growth. *Quart J Econ* 70(1):65–94
- Swan TW (1956) Economic growth and capital accumulation. *Econ Record* 32(2):334–361

Endogenous Growth Theory and Regional Extensions

11

Zoltan Acs and Mark Sanders

Contents

11.1	Introduction	194
11.2	Modern Growth Theory	195
11.2.1	Empirical Findings on Growth	195
11.2.2	Modeling R&D as the Source of Growth	196
11.2.3	Modeling Education as the Source of Growth	198
11.2.4	Concluding Remarks	199
11.3	The Geography of Knowledge Creation and Diffusion	199
11.3.1	Clusters and Agglomeration	199
11.3.2	Knowledge Spillovers and the Role of Cities	203
11.3.3	Concluding Remarks	204
11.4	The Organization of Knowledge Creation and Diffusion	205
11.4.1	Organizational Change as the Enabler of Economic Growth	205
11.4.2	Entrepreneurship as the Conduit for Knowledge Spillovers	205
11.4.3	Concluding Remarks	208
11.5	Fundamental Causes of Growth and Development	208
11.5.1	Institutions and Economic Growth	208
11.5.2	Concluding Remarks	209
11.6	Conclusions	210
	References	210

Z. Acs (✉)

School of Public Policy, George Mason University, Fairfax, VA, USA
e-mail: zacs@gmu.edu

M. Sanders

Utrecht School of Economics, Utrecht, The Netherlands
e-mail: M.W.J.L.Sanders@uu.nl

Abstract

In this chapter, we outline the basic mechanisms in endogenous growth theory that identify knowledge creation and diffusion as the core driver of economic growth. Then we discuss how new economic geography, urban economics, organizational science, and entrepreneurship theory have *regionalized* the mechanisms involved. Knowledge creation, however, has been dubbed a *proximate cause* of growth, and the quest for *fundamental causes* has continued. We then discuss this recent development in macroeconomic growth theory and argue that the new *institutional approach to growth* opens up a lot of new avenues for further research. Once again, the importance of cities, organizations, and entrepreneurship is being ignored in macro growth theory. Yet economic geography, urban economics, organizational science, and entrepreneurship theory have a lot to contribute to growth theory by both empirically and theoretically developing our understanding of local institutions and linking these to regional economic development and growth.

11.1 Introduction

Growth at the regional level can be understood as the increase in economic activity in that region over a period of time. We can measure this empirically (with all well-known caveats) by the annual increase in GDP or employment at the regional level. The sources of economic growth have always been the subject of intense academic investigation. However, the resulting state-of-the-art modern economic growth theories – for example, presented in Barro and Sala-i-Martin (2004) and Acemoglu (2009) – largely abstract from geography. That is, the process of economic growth is investigated without considering the location and regional setting in which it is taking place. Of course, that will not do for the purposes of this handbook.

The literature shows that there is large variation in regional levels and growth of GDP. Barro and Sala-i-Martin (2004) present tests of convergence among regions, showing slow convergence exists, but migration and capital mobility fail to eliminate regional variation in development levels, even over long periods of time and among well-integrated regions. We have seen many possible explanations being offered and empirically investigated in other chapters of this part of the handbook. All in one way or another suggest that although cross-regional spillovers are significant, the growth and development process is also very much a localized process. In this chapter, we will discuss at some length the models and theories that explicitly link economic activity and development to regional circumstances.

But before we discuss such extensions to the basic framework, it is useful to review modern growth theory. Then we can discuss the various ways in which these models have been *regionalized* in the literature on new economic geography, urban economics, and the spillover theory of entrepreneurship. These literatures all build on the first generation of ideas in innovation-driven endogenous growth theory. We can then review the implications of the more recent institutional approach to

economic growth. Barro (1996) and Acemoglu, Johnson, and Robinson (2001) pioneered this approach, but here too, geography is largely ignored. One way to link institutions and growth at the regional level is to consider the link Baumol (1990) proposed between productive entrepreneurship and institutions. The institutional approach to growth may well open up new and exciting research avenues for building regional endogenous growth models. As this literature is very much in an embryonic stage, however, the final sections will raise questions for further research.

11.2 Modern Growth Theory

11.2.1 Empirical Findings on Growth

In his analysis of changes in the aggregate production function, Solow (1957) referred to earlier research that found that over 90 % of the variation across US states in GDP per capita growth was due to productivity increases. His work sparked a large literature on the rate of technological change, where the empirical literature sought to adequately measure productivity and look for statistically significant correlations with possible explanatory variables (e.g., Denison 1967). The (mainstream) theoretical literature, meanwhile, built on the evidence and formulated new hypotheses and models that try to endogenously explain technological change and productivity growth as the result of (rational) behavior of economic agents.

At around the same time as Solow, however, Kaldor (1963) posited his stylized facts of growth and presented economists with another puzzle. Kaldor (1963) showed that over time, while per capita capital stocks grow, the income shares in GDP are approximately stable. This implies that technological change is biased and purely labor augmenting. This finding sparked a literature on the bias in technical change that received a big push from the observation that in the 1980s the labor income share of the low skilled workers started dropping dramatically. The empirical literature again focused on measuring and separating the various biases in technological change and related them statistically to a host of explanatory variables while the theoretical literature struggled to develop a model that would explain how rational agents generate such biases in response to price and demand developments.

As innovation is an illusive concept that is hard to measure or observe directly, the empirical evidence in support of the claim that innovation is important is convincing yet circumstantial. Technological change and innovation have been identified as the source of long-run economic growth by elimination of all other possible sources. As more and better data became available, recent studies then found regularities in the data that strongly suggested that the rate of innovation indeed drives long-run economic growth. The empirical literature on economic growth has been surveyed in much more detail than we can hope to achieve in this chapter by, for example, Temple (1999) and Barro and Sala-i-Martin (2004). From their surveys, we learn that growth rates tend to remain positive in the long run and differ a lot for long time spans among countries. Temple (1999) surveys the

evidence that links these cross-country differences to the proximate causes of growth: investment in physical capital (machinery, equipment, and infrastructure), investment in human capital, and investment in research and development.

In particular, in explaining differences between developing and developed countries, the investment in physical capital seems a strong candidate. Differences among developed countries are attributed *inter alia* to knowledge creation (e.g., R&D), although there is also strong evidence for international spillover of knowledge. The evidence on human capital accumulation through education is mixed. The empirical literature on these issues, however, typically struggles with the statistical problem to separate cause and effect as investments in knowledge creation and education may both affect and be affected by economic development.

In addition to these proximate causes, Temple (1999) discusses evidence on the impact of population growth, trade openness, financial development, macroeconomic stability (low and stable inflation and unemployment rates), inequality, political and civil rights, the size of government, and public infrastructure investments. All these factors are, however, believed to work through, or as moderators of, the impact of knowledge creation.

In the empirical literature, one can find various exogenous and semi-endogenous representations of technological change. Through introducing time trends, accumulated production, new capital investment, imports and exports, etc., as arguments in growth regressions or regression equations for production frontiers or cost curves, this literature has explicitly or implicitly introduced and developed ways to represent learning by doing (cumulative production), learning by investing (capital accumulation), embodied technical change (new capital goods), imported technology (imports), and learning by exporting (exports). The latter representations, however, can still not be called *endogenous technological change* in the true sense of the word. The empirical models assume a relation between technology and other (endogenous) variables such as cumulative production, but do not explain the mechanism and behavior behind this relation. But this is as far as empirics can take us. Going from empirical measures and representations of technology to actually explaining the phenomenon requires a theoretical model that illustrates how behavior of agents in response to incentives and constraints leads to innovation and technological change.

A fundamental proposition in new growth theory is that rational agents create and accumulate knowledge for a purpose. In accordance with the findings in the empirical literature discussed above, theory has explored R&D and education (human capital) as the sources of growth. In addition, and more recently, organizational change and institutional development have been identified as key enablers or *fundamental causes* of economic growth.

11.2.2 Modeling R&D as the Source of Growth

With Romer (1990), Aghion and Howitt (1991), and Grossman and Helpman (1991a), a strand of models has developed in which R&D, the conscious and

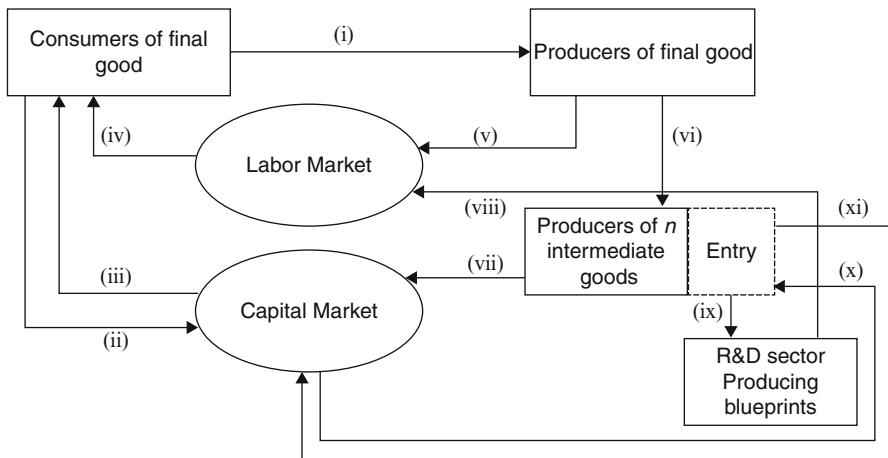


Fig. 11.1 The Romer (1990) model

deliberate investment of resources in knowledge creation, generates inventions that generate productivity growth. The structure of the Romer (1990) models is illustrated in Fig. 11.1.

The arrows (i)–(xiii) represent financial flows. Consumers spend on final goods (i) and save part of their income (ii) by purchasing assets. Their income consists of capital (iii) and labor income (iv). The final goods producers hire labor (v) and buy intermediate goods (vi) that are imperfect substitutes in production and supplied by monopolistic intermediate producers that only hire raw capital (vii) as their input. Given that there are monopoly profits in intermediate goods production and latent demand for new intermediates is positive, there is an incentive for the R&D sector to hire labor (viii) and produce new blueprints that are sold (ix) to new entrants, who have to borrow (x) and repay (xi) this loan. The diagrams for the Grossman and Helpman (1991a) (there are no intermediates, but final goods are imperfect substitutes in consumption) and Aghion and Howitt (1991) (there is no variety expansion but quality improvement of existing intermediates) models would look slightly different, but as Grossman and Helpman (1991a) show, the qualitative results and key driving mechanisms remain the same. Based on these four basic models, a large literature has emerged, presenting models that focus on the accumulation of tradable knowledge embodied and codified in patents, products, and processes.

The basic behavioral assumption in these models is perfect rationality. Representative agents maximize utility or profits and are constrained by market (demand functions) and technological (production functions) circumstances. These behavioral assumptions are very strong and lack a strong empirical basis. The models should therefore not be taken too literally. Rather, in explaining the mechanism through which an economy with pure rational agents would generate innovation, the models show why it pays to innovate. Such mechanisms then also produce a tendency toward productivity growth when more realistic behavioral heuristics are applied. Such extensions and adaptations, however, usually come at the cost of

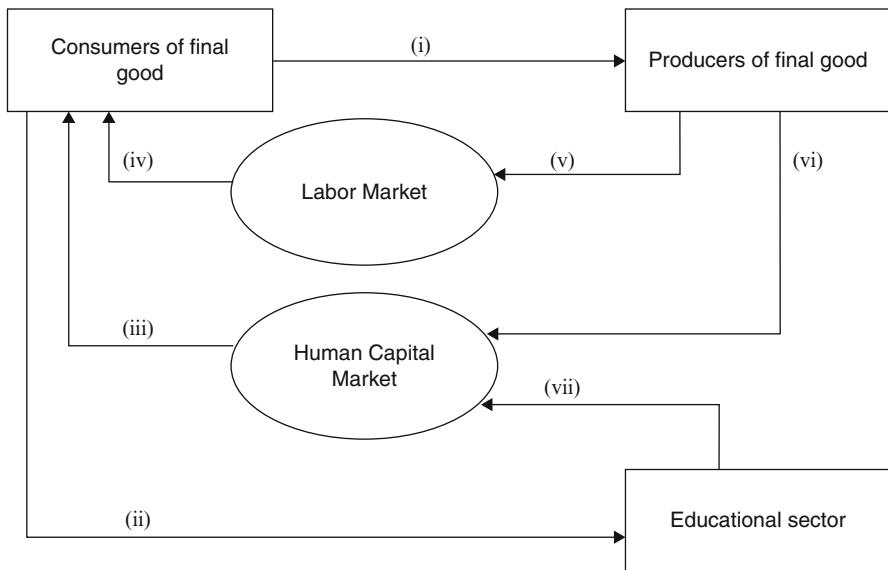


Fig. 11.2 The Lucas (1988) model

more complexity and loss of tractability. As consumption, (intermediate) production, and R&D are typically geographically concentrated and distributed in space, a clear link to regional models of growth can easily be built in.

The key insight to be taken away from these models is that investments in knowledge creation will generate innovation and economic growth when the private returns to knowledge creation, ultimately related to the consumers' preferences, justify private investments in knowledge creation. The public good nature of knowledge (non-rivalry) then implies increasing returns to knowledge accumulation in the aggregate and allows for growth to be positive in equilibrium. Typically, the knowledge generated in the past is assumed to be available and increase the productivity of knowledge generation in the present. This externality drives growth and is assumed to arise automatically, but it is not a big leap to link such spillovers to regional and local variables such as proximity and density. The private rewards to innovation, in addition, typically come from some sort of entry barrier that protects the innovator. This is another way in which (local, regional, or (inter)national) conditions enter the debate.

11.2.3 Modeling Education as the Source of Growth

Following the original ideas of Becker (1964), Lucas (1988) pursued the idea that human capital accumulation through education generates productivity growth. Lucas (1988) presented a model (see Fig. 11.2) in which rational agents consume (i) and invest in their education (ii) in response to expected private future benefits. Production uses raw labor (iii) and human capital (iv), so it benefits from higher human

capital levels, but because the educational sector also uses human capital (v), education makes both production and the future education process more productive in his model. Lucas thereby introduces aggregate increasing returns to knowledge accumulation as in the models above. In this model, however, the knowledge accumulates in the working population, not in some disembodied knowledge base or in the technology embodied in the physical capital stock. It would therefore make sense to interpret that type of technological change as quality improvement of labor inputs.

The key to generating endogenous growth with human capital is again introducing externalities and spillovers that create increasing returns at the aggregate level. As Lucas (1988) himself immediately recognized, however, the productivity of human capital depends a lot on the presence of other educated workers in close proximity. He therefore discussed at some length the role of cities in economic growth and of human capital spillovers in holding cities together.

Of course, the educational level of the population also plays a role in the R&D-driven growth models, as R&D typically requires high-skilled (technical) personnel. The two models combined would therefore allow for education to affect productivity directly but also indirectly through the increased (productivity of) R&D activities.

11.2.4 Concluding Remarks

Modern economic growth theory has focused on knowledge accumulation in the workforce and the aggregate production system as the long-run sources of economic growth. It models how rational decision makers are privately induced to produce knowledge and thereby economic growth as a positive externality. Education makes all workers more productive, new technologies make further innovation easier, and knowledge, once created, is non-rival in use. Economic growth comes from the increasing returns to human capital and knowledge creation that exist at the aggregate level.

The first generation of endogenous growth models thus predicted that economies with a larger workforce would experience higher growth (the so-called scale effect), which flew in the face of empirical evidence. This led many to propose alternative specifications that did not suffer from this apparent flaw. But once one realizes that the *aggregate level* of an economy is not the same as the size of a nation's workforce and the appropriate level of analysis is not the national but the regional or even local level, scale and density become similar concepts, and we can start to rethink the implications of endogenous growth in a slightly different direction.

11.3 The Geography of Knowledge Creation and Diffusion

11.3.1 Clusters and Agglomeration

It is a well-established empirical regularity in economic geography that population, and with it economic activity, is highly concentrated in urban centers. In addition, it follows

often at least approximately a remarkably stable distributional pattern, known as Zipf's law. Zipf's law states that population centers, ranked by size, will decline in size proportionately to their rank. This implies that if the second city in a region has half the population as the first, the third has one third, etc. Moreover, Zipf's law has been shown to hold for the distribution of population in space at the national, regional, and even agglomeration level in, for example, Germany by Giesen and Suedekum (2011).

This pattern of spatial distribution interacts with another important empirical regularity in economic geography, known as the gravity equation. The gravity equation was proposed to explain the intensity of interactions between countries but can also be applied to interaction between and within clusters of economic activity at lower aggregation levels. The gravity equation explains, for example, trade and migration flows as a function of GDP of the origin and destination regions and the distance between two locations. In empirical work on Zipf's law and the gravity equation, the explanatory power of these relatively simple models is often above 90 % of the variation to be explained across regions. This suggests that core-periphery clusters are remarkably common, and their general spatial structure and the linkages between constituent parts are similar.

Early work on the economics of agglomeration of industries identified the availability of inputs (specialized labor, other non-tradable specialized inputs) and access to output markets as key reasons for the spatial agglomeration of industries and argued that in agglomerations, the informational spillovers would improve productivity. In his seminal contribution, Paul Krugman (1991) presented a model that would generalize this intuition using among other things newly developed endogenous growth modeling tools. By now, there are many excellent textbook treatments of his basic model.

Intuitively, the model exploits a *circularity* or positive feedback loop. Consider an economy with agriculture and manufacturing. Agriculture is obviously located where the land is fertile. With high transportation costs and limited economies of scale in manufacturing, manufacturers will want to locate close to their demand. In a largely agricultural economy, this implies manufacturing is distributed in space pretty much in the same way as arable land. If on the other hand economies of scale are strong and transportation costs are low, manufacturing will begin to cluster. And if manufacturing workers also provide most of the demand for manufactured products, the process of agglomeration will feed on itself.

The model can be represented in a figure similar to the one provided for the Romer (1990) and Lucas (1988) models above. It is immediately clear from Fig. 11.3 that the ingredients of the Krugman (1991) model can easily be connected to the core elements in innovation-driven endogenous growth models. Krugman (1991) assumed two regions and imperfect substitutes and monopolistic competition in both manufacturing sectors. That is, consumers from both regions receive labor income (i) and consume local food (ii), but both purchase locally produced (iii) and imported (iv) manufactures. Both food production (v) and manufacturing (vi) use labor only in the basic model. Krugman (1991) then focused on labor migration and investigated how differences in wages (vii) might trigger movements of labor, explaining the emergence of core-periphery patterns.

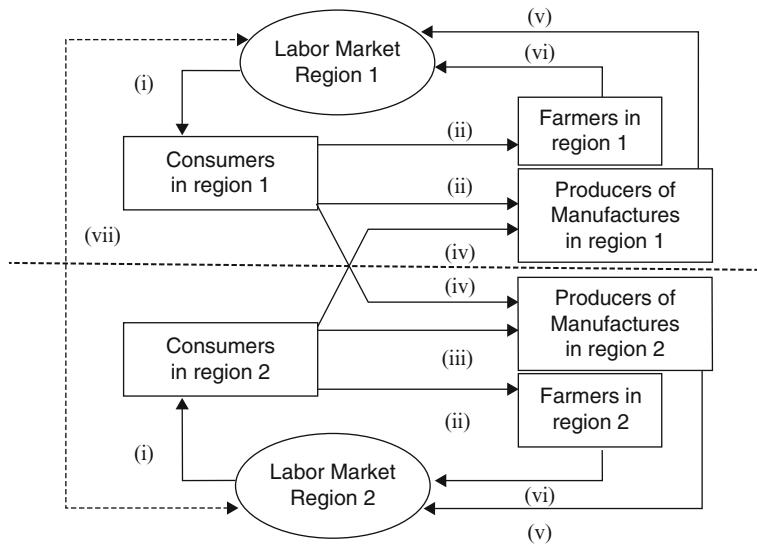


Fig. 11.3 Krugman (1991) model

The Krugman (1991) model predicts that in a perfectly symmetric situation, the wages for manufacturing workers are equal in both regions, and there will be no pressure for migration or clustering. If, however, the initial manufacturing workforce is not symmetrically distributed, there are three competing effects on the relative (real) wage of manufacturing workers. On the one hand, the region with the larger market pays higher manufacturing wages because scale economies reduce costs and the larger market is cheaper to service in the presence of transportation costs. In addition, the lower price of manufactures due to competition increases the real value of wages in the larger market. On the other hand, there is less competition for the demand for manufactures from agricultural workers in the smaller market. The strength of these competing forces depends on the share of agriculture and manufacturing in total income, the degree of competition among manufacturers, and the level of transportation costs.

For regional endogenous growth, this model has important implications. The sources of aggregate increasing returns in modern manufacturing and services sectors are easily related to the sources of increasing returns that drive endogenous growth in modern growth theory. Baldwin and Forslid (2000), for example, introduced a Romer (1990) R&D sector into the Krugman (1991) model by allowing for new manufactured goods and show that the integration of knowledge bases creates growth, even in the periphery, that compensates for losses of agglomeration there. What these models have in common is that they make regional development highly path dependent. Initial conditions determine to a large extent where what type of economic activity will establish itself and how it will (fail to) grow.

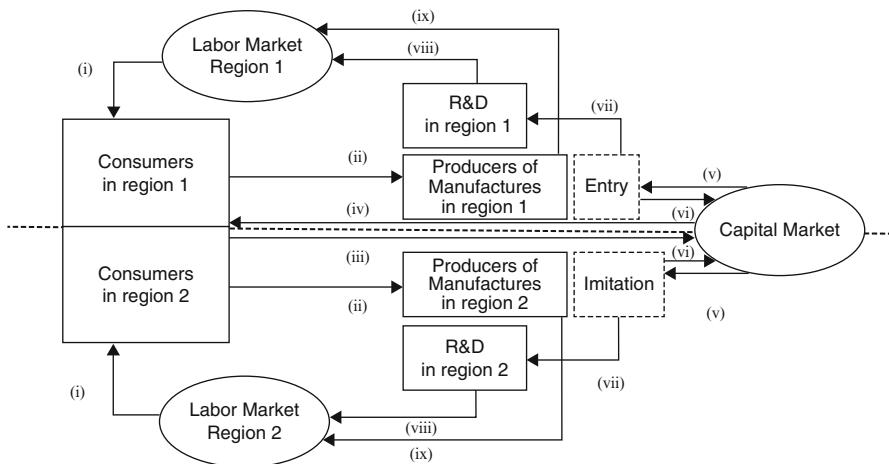


Fig. 11.4 The Grossman and Helpman (1991a) model

A model that explicitly uses endogenous growth theory is a model of the product life cycle by Grossman and Helpman (1991b). As can be seen in Fig. 11.4, in this model, Grossman and Helpman introduce an R&D sector in both regions but have the advanced region develop new products, whereas the lagging region is merely imitating them.

In the model, total wage income (i) is spent on consumption (ii) of diversified final goods from regions 1 and 2. Consumers also save in a global capital market (iii) and receive interest (iv) on their assets. These savings are invested in new ventures (v) and repaid out of profits (vi). The required knowledge for a new or imitation venture is produced by and auctioned off by a local R&D sector (vii), and both the R&D and production sector compete and pay for labor (viii) and (ix), closing the model. There is free trade, but all final goods production is monopolistic, and consumers value variety and new product designs. This implies monopoly rents are available to incentivize knowledge creation. The backward region in the model, however, does not invent but rather copies the products developed in the advanced region, and of course, this is easier when a lot of advanced products are available. Thus, the model explicitly endogenizes regional and intertemporal knowledge spillovers and shows that growth in the periphery will benefit from knowledge creation in the core. In fact, in equilibrium, the growth rates in both regions will equalize, and the core has to *run to stand still* in terms of relative wages and income levels. Audretsch and Sanders (2011) extended this model to a three-stage life cycle, showing *inter alia* that new labor-rich countries and regions joining the global economy will leave the advanced region a more limited range of products and cause the fringe to catch up. In these models, however, information and knowledge will flow between the *advanced* and the *backward* region as water will run down a hill. Moreover, this model assumes labor immobility, suggesting the regions are geographically not very near.

11.3.2 Knowledge Spillovers and the Role of Cities

As endogenous growth theory has emphasized knowledge accumulation, either in the labor force or in the form of new products and processes, it is useful to distinguish *knowledge* from *information* as in Feldman and Audretsch (1999). Information is codified and can be transmitted over large distances at low costs. In fact, with modern communications technology, the costs of reproducing and transferring information are effectively zero. Knowledge, in contrast, is partially tacit and requires proximity and high-quality two-way communication for effective knowledge transfer. If the positive externalities and increasing returns depend on others also benefiting from past knowledge creation, the intensity of personal interaction becomes a relevant factor in determining growth to the extent that knowledge is not information. And as this intensity is higher and more diverse in a densely populated region, this puts urban, high-density regions center stage in the discussion.

Glaeser et al. (1992) in their seminal study of growth in cities tested three alternative theories of endogenous knowledge-spillover-driven economic growth in a sample of US cities. They tested the Marshall-Arrow-Romer approach that relies on within industry knowledge spillovers and monopoly profits to motivate knowledge creation against the Porter approach that also emphasizes intraindustry knowledge spillovers but rather relies on competitive pressures to innovate and the Jacobs approach that instead relies on interindustry knowledge spillovers, where ideas cross-fertilize over industry boundaries. Perhaps surprisingly, Glaeser et al. (1992) found that the data favor the latter theory, suggesting that specialization at the regional level is detrimental for growth. Feldman and Audretsch (1999), however, found that specialization does promote regional growth and development. De Groot, Poot, and Smit (2009) concluded from a meta-analysis of this literature that although most research supports Porter and Jacobs, the heterogeneity across regions, periods, and sectors seems large.

Whether good or bad for growth, it is an established fact that cities and urban regions typically are quite specialized and have become increasingly so. Smaller cities are more specialized than the large metropolitan areas, however, so as cities grow in size, they also tend to become more diversified. This is because larger cities attract the much more diversified business services. So cities specialize, and so do industries. Ellison and Glaeser (1999) studied the reason for industries to concentrate in regional clusters and found that only 10–20 % of the spatial pattern could be explained by natural advantages (access to port facilities and navigable rivers, presence of natural resources, etc.). The remaining 80–90 % would have to be attributed to intraindustry spillovers. Cities attract these so-called *footloose* production activities, where location is not important, but agglomeration is.

Florida (2005) discussed the impact of the creative class on regional development and finds that more open cultures perform better. This suggests that it is indeed the exchange of knowledge, particularly tacit knowledge, that drives prosperity and growth. And the exchange of tacit knowledge requires open and intensive communications between people of very diverse backgrounds and perspectives.

Only considering the positive knowledge spillover externalities in the models, however, would lead us to the conclusion that agglomerations are typically too small. Given that there are positive externalities connected to the decision to locate in a given area by increasing its density, it would then make economic sense for cities to grow without bounds. But of course, there are also negative externalities to consider. Henderson (2006) presents models of urban growth that rely heavily on endogenous growth models as described above. But he explicitly introduces the negative externalities connected to congestion and commuting. This introduces an inverted u-shaped relationship between city size and total welfare and implies cities have an optimal size. This optimal size is shown to depend negatively on transportation and congestion costs and positively on benefits of agglomeration, total factor productivity, and human capital accumulation. Henderson shows that it depends on their institutional setting and quality whether actual cities will achieve this optimal size. One needs to either have local governments that can limit city size or free markets in which both negative and positive externalities are priced into land rents and local public goods provision, respectively. In the absence of such institutions, as arguably is the case in many developing countries, cities will grow beyond their optimal size. When the marginal migrant is indifferent between living in an overcrowded city or a sparsely populated rural village, the positive externalities of density in knowledge creation are fully offset.

11.3.3 Concluding Remarks

This section has reviewed the main models that turn endogenous growth models into models of regional endogenous growth. As endogenous growth models build on positive externalities connected to knowledge accumulation at the aggregate level, regional models of endogenous growth all have in common that they localize this process. New economic geography models will distinguish between knowledge accumulation in agriculture, manufacturing, and services to explain the core-periphery patterns of economic activity, whereas urban economics models explain the existence and growth of urban agglomerations by focusing on the centripetal forces of localized knowledge spillover externalities. These models thus explicitly localize the knowledge spillovers that cause increasing returns in the aggregate growth models discussed above. One can localize the knowledge base on which new knowledge is created or link the creation of positive spillovers to density and proximity to explain spatial patterns of economic development. What these models do not endogenize, however, is the selection of what agents will take what actions to create and/or diffuse the knowledge into the regional economy. New economic geography and urban economics share with endogenous growth theory a relative neglect for the organizations and people that create or build on the new knowledge.

11.4 The Organization of Knowledge Creation and Diffusion

11.4.1 Organizational Change as the Enabler of Economic Growth

Aghion et al. (1999) pointed out the possible relevance of organizational change for economic growth. The organizational change literature stresses the fact that introducing new technology always requires a rethinking of the organization of the production process. In this literature, however, organizational change is not considered a root cause of economic growth and innovation. By optimizing the organization, productivity can be increased for a given set of technologies. But it is likely that this type of productivity increases run into severe diminishing returns in the absence of true technological innovation or knowledge accumulation.

Organizational change can never improve efficiency above 100 % and push the technology frontier out by merely reorganizing the existing production techniques, knowledge, and outputs. The rate of knowledge accumulation thus ultimately limits productivity growth. But only when the appropriate organizational changes are made will newly acquired knowledge actually create growth.

For the current chapter, this model relies primarily on the short- and medium-run adjustment process to new technology. Adopting new technology requires adaptive skills at the regional, local, and organizational level. The absence of such skills will slow down the rate of technological change even if knowledge creation is not the bottleneck in the innovation process, whereas an abundance will facilitate spillovers from outside. The Aghion et al. (1999) approach looks at this adjustment to new knowledge and technology as taking place predominantly *within* existing organizations and concludes that generally high-skilled workers will be in higher demand when technical and therefore organizational change is intense. The implicit assumption in this framework is that existing firms and organizations adopt and commercialize new knowledge and ideas. The entrepreneurship literature has focused instead on new firms and organizations as the vehicle for new knowledge commercialization.

11.4.2 Entrepreneurship as the Conduit for Knowledge Spillovers

Endogenous growth theory has largely considered knowledge accumulation as the key determinant of growth and consequently focused on explaining the accumulation of new knowledge. New economic geography and urban economics had already recognized the importance of local knowledge spillovers and were quick to link the tools in endogenous growth theory to local and regional economic activity. And organizational change theory addressed the skill requirements that follow from knowledge implementation in existing organizations. All these approaches, however, ignored the Schumpeterian distinction between *invention*, the creation of an idea, and *innovation*, the commercialization of that idea. The working assumption in most of the models discussed above is that inventions will

automatically or trivially diffuse throughout the (local) economy causing innovation. There are telling examples, however, of firms that failed to commercialize the new knowledge developed in their own R&D departments. Similarly, regions and organizations have sometimes failed to develop competitive advantages based on knowledge that was available. There can be many reasons, but it is obvious that idle knowledge will not cause economic growth. More generally, knowledge creation is a fundamental and definitely necessary, but not a sufficient, condition for generating economic growth.

A large body of literature on entrepreneurship shows that innovation and knowledge spillovers are far from trivial and automatic. Moreover, the process of commercialization is driven and restricted by very different incentives and constraints than knowledge creation. Arguably, the entrepreneurial skill, attitudes, and infrastructure in a region will then co-determine how effectively that regional or urban economy implements and benefits from knowledge creation. There is quite some empirical evidence that supports the claim that at regional and local levels, it is the (lack of ambitious) entrepreneurship that explains the success and failure to turn knowledge creation into economic growth. The entrepreneurship literature long took the creation and existence of new business opportunities as exogenous and zoomed in on the entrepreneurial process of recognizing and exploiting opportunities instead. Its focus to date is largely empirical and shows strong regional differences in entrepreneurial activity that are closely correlated with economic growth and development. These differences can of course be related to the local presence or absence of new knowledge and ideas, but it has been shown that many other regional and individual variables play a role. Entrepreneurial attitudes and ambitions, local outside options for entrepreneurs and the presence of many small firms and entrepreneurs, a skilled labor force, and a high population density all seem to be of importance.

In Acs et al. (2009) and Acs and Sanders (2012), it was argued that entrepreneurship should be considered the *missing link*, the key conduit (and bottleneck) for knowledge spillovers. It is not through the mere presence of or random interaction among many agents in a densely populated urban setting that knowledge diffuses and creates the positive externalities that endogenous growth theory assumes. The *knowledge spillover theory of entrepreneurship* (KSTE) as Acs et al. (2009) dubbed their approach argues that it is entrepreneurs that deliberately transfer knowledge and transform invention into innovation in the pursuit of profit. In doing so, this theory endogenizes the spillovers that were assumed to exist or arise costless and semi-endogenously in endogenous growth models discussed above. In that sense *an entrepreneurship theory of knowledge spillovers* might have been equally appropriate.

As such, the KSTE opened up a range of new opportunities for thinking about endogenous models of regional growth. Consider the Acs and Sanders (2012) model represented in Fig. 11.5, where we took the Romer (1990) model and added the possibility for the final goods sector to hire labor to do R&D (vii). To fund these investments, they need to borrow (viii) and repay (ix) their loans from

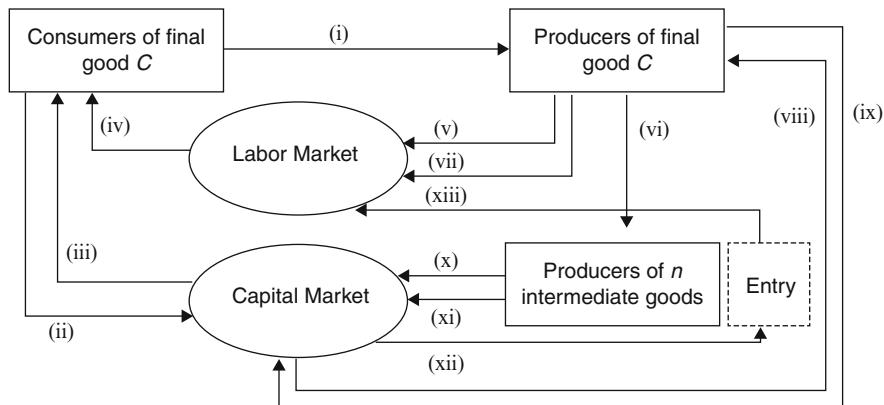


Fig. 11.5 The Acs and Sanders (2012) model

additional profits. The R&D in the final sector creates *upstream* knowledge spillovers that are free. The R&D in a car-manufacturing firm, for example, may create an occasional opportunity for an entrepreneur to become a new supplier of some part in the intermediates sector. The entrepreneurs can set up such a firm, but this requires labor (xiii) that has to be financed (xii), implying all existing intermediate firms also repay the start-up loan (xi).

The implications of the model are intuitive. In the absence of a pool of ambitious and skilled entrepreneurs, the knowledge spillovers are limited to within firm accumulation of knowledge. Entrepreneurs facilitate the knowledge spillovers that seem important in the Jacobs and Porter approaches, described above. One can also put more emphasis on the occupational choice trade-offs, implicit in the model. By making entrepreneurship and employment perfect substitutes, wages equalize and one implicitly assumes that workers are indifferent and homogenous. But of course, one can extend this simple model in the spirit of Krugman (1991) and include many other individual and regional variables that will affect the basic trade off. Such a model would explain clustering of entrepreneurs in urban regions and detach the dynamics in the global knowledge base from local knowledge implementation. In dynamic settings with multiple regions, one will then be in a steady state when the labor allocation is more or less stable across regions and occupations and the marginal products of labor in knowledge generation, knowledge commercialization, and routine production equalize.

The many different externalities that exist in all three stages of the innovation life cycle all provide a link to location and urban economics, as some are closely related to proximity and density and others less so. The key conclusion from this theory is, however, that more knowledge creation will increase growth and development only through the innovative activities of entrepreneurs who turn information and ideas into a widely used and readily available knowledge base that facilitates further knowledge creation.

11.4.3 Concluding Remarks

When we turn our focus on the organizations and people that create and commercialize the new knowledge, new linkages through which endogenous growth is localized can be explored. On the one hand, existing organizations require skilled employees to absorb appropriate new technologies from outside and from within. In the absence of such absorptive capacity, existing firms in a region will not be able to internalize and build upon the knowledge created. The spillover theory of entrepreneurship on the other hand stresses the role of entrepreneurs in transferring knowledge from the drawing board to the market place, providing an alternative missing link between invention and innovation. Both approaches shift the focus from local knowledge creation to local knowledge commercialization, where different mechanisms may be at work.

11.5 Fundamental Causes of Growth and Development

11.5.1 Institutions and Economic Growth

Institutions are broadly defined as *the man-made rules of the game*. For a long time, (growth) economists have treated institutional differences, much like, for example, differences in geography and preferences, simply as given. In searching to explain the cross-country differences in growth performance over longer periods of time, however, several authors have now argued that institutions play an important role in determining growth. The empirical evidence that causality runs from institutional quality to economic performance is mounting, since Acemoglu, Johnson, and Robinson (2001) showed a clear link between sixteenth-century settler mortality rates and modern institutional quality and used this as an instrument in their empirical analysis. Without going into technical detail too much, this chapter clearly establishes the importance of institutional quality for development and growth. This takes the endogenization of economic growth and innovation one step further. The institutional approach is more fundamental than the modeling approaches above and puts a strong emphasis on growth as a historical process.

Still, the institutional approach to date largely abstracts from geography and location. Acemoglu (2009) dismisses geography (and the alternative hypotheses that it is all about luck or culture) as a fundamental cause of economic growth by arguing that climatic and geographic circumstances fail to explain why some countries did and some did not industrialize. This exclusive focus on *first-nature* geography obscures the importance of second-nature geography that new economic geographers and urban economists tend to stress and that has been the core of their work. It is, however, hard to see where this neglect will take institutional growth theory. Recent contributions focus on explaining the endogenous emergence and persistence of growth-enhancing or inhibiting institutions, where the natural inclination is to think about political institutions at the national level supporting or inhibiting knowledge creation. Models in this field, for example, explain the

transition from autocracy to democracy as a result of power struggles and apply game theory and political economy techniques to explain institutional change. Many of the formal institutions (property rights, rule of law, democracy, etc.), however, do not vary across regions and locations in the same country, whereas informal institutions (culture, attitudes toward the new, experimentation, etc.) differ quite a bit between core and periphery in general and across different regions and cities as well.

As Florida (2005) has shown, more open cultural attitudes also correlate highly with standards of living and economic performance. And Baumol (1990) already argued that it is institutions that will also mobilize entrepreneurial talent into or out of the productive activity of taking knowledge from the drawing board to the market place. Both would probably agree that it is local, informal institutions in cities and regions that matter as much or more than the formal, legal, and political institutions at the national level for enabling and promoting the knowledge spillovers that endogenous growth theory assumes at the aggregate level.

An interesting but to date largely unexplored line of research would therefore focus more on the importance of local and regional institutional differences and zoom in on their beneficial or detrimental effects on knowledge creation and diffusion at the regional level. Attempts in this direction have been made by scholars taking a much more empirical and case-based *regional systems of innovation* approach that is very popular with policy makers that try to identify regional strengths and weaknesses, also in the institutional setting. Anselin et al. (1997) already showed that knowledge spillovers are highly localized, whereas Audretsch and Lehman (2005) establish a link between local knowledge creation and commercialization activity.

The institutional approach to economic growth is currently asking very different questions, however. Inspired by aggregate endogenous growth theory, it focuses almost exclusively on the process of knowledge creation, and the supporting empirical work will continue to draw predominantly on panel databases that have countries as their geographic unit.

11.5.2 Concluding Remarks

Recent contributions to the empirics of economic growth have established that institutions play a key role in driving growth and development at the aggregate level. The current challenge in this literature is to link these national institutions to the process of knowledge creation. We feel that such an effort is likely to suffer from the current neglect of endogenous knowledge spillovers at the aggregate level. Likewise, recent evidence in economic geography suggests that local growth and development patterns are driven largely by the interaction between knowledge creation and commercialization, but this literature still largely ignores local institutions as, for example, very different cultural attitudes toward entrepreneurship and innovation in cities and their periphery. The institutional approach to aggregate economic growth will remain free of geography as long as it ignores the importance

of proximity for the assumed knowledge spillovers, whereas economic geography and urban economics will remain free of institutions as long as their importance for attracting and mobilizing creative and entrepreneurial talents is not recognized. Of course, exploring the implications of institutional differences among regions and cities on the one hand and second-nature geography in knowledge creation and spillover opens up a broad research agenda for both theory and empirical research in the field of endogenous regional growth.

11.6 Conclusions

In this chapter, we have first reviewed basic endogenous growth theory to show that economic growth is ultimately dependent on new knowledge creation. This new knowledge creation, however, builds on the existing knowledge base, causing increasing returns to scale at the aggregate level. This same mechanism has been shown to operate at the local and regional level in models of new economic geography and urban economics that aim to explain spatial patterns of economic activity and growth. More specifically, we showed that the mechanism of endogenous economic growth can account for stable city growth and core-periphery patterns of economic activity. When we dig into the process of knowledge creation a bit more, we have to distinguish between knowledge creation and commercialization. The former is of course a necessary condition for growth, but the latter may well be the more important bottleneck for growth at the regional level. In addition, commercialization is a way to make new knowledge available to those who would otherwise not be able to build upon it. Arguably, the commercialization, rather than the inception of a new idea, has the biggest impact on economic performance and enables future knowledge spillovers to take place. Moreover, a focus on the people and organizations that are responsible for commercialization allows for another fruitful avenue for localizing endogenous growth. Finally, we have discussed how the most recent developments in economic growth theory, toward a more institutional approach to economic growth, once more seem to neglect the importance of geography. Institutions have been shown to affect local, organizational absorptive capacity and both the attitudes and actions of entrepreneurs. Linking such insights with the modeling tools as they are developed in the mainstream of macroeconomic growth theory will certainly constitute a fruitful and important agenda for future research.

References

- Acemoglu D (2009) Introduction to modern economic growth. Princeton University Press, New York
- Acemoglu D, Johnson S, Robinson J (2001) The colonial origins of comparative development: an empirical investigation. *Am Econ Rev* 91(5):1369–1401
- Acs Z, Sanders M (2012) Patents, knowledge spillovers and entrepreneurship. *Small Bus Econ* 34(4):801–817

- Acs Z, Audretsch D, Braunerhjelm P, Carlsson B (2009) The knowledge spillover theory of entrepreneurship. *Small Bus Econ* 32(1):15–30
- Aghion P, Howitt P (1991) A model of growth through creative destruction. *Econometrica* 60(2):323–351
- Aghion P, Caroli E, Garcia-Penalosa C (1999) Inequality and economic growth: the perspective of new growth theories. *J Econ Lit* 37(4):1615–1660
- Anselin L, Varga A, Acs Z (1997) Local geographic spillovers between university research and high technology innovations. *J Urban Econ* 42(3):422–448
- Audretsch D, Lehmann E (2005) Does the knowledge spillover theory of entrepreneurship hold for regions? *Res Policy* 34(8):1191–1202
- Audretsch D, Sanders M (2011) Technological innovation, entrepreneurship and development. In: Szirmai E, Naude W, Goedhuys M (eds) *Entrepreneurship, innovation and economic development*. Oxford University Press, Oxford, pp 35–64
- Baldwin R, Forslid R (2000) The core-periphery model and endogenous growth: stabilizing and destabilizing integration. *Economica* 67(267):307–324
- Barro R (1996) Democracy and growth. *J Econ Growth* 1(1):1–27
- Barro R, Sala-i-Martin X (2004) *Economic growth*, 2nd edn. The MIT Press, Cambridge
- Baumol W (1990) Entrepreneurship: productive, unproductive and destructive. *J Polit Econ* 98(5):893–921
- Becker G (1964) *Human capital*. Columbia University Press, New York
- Denison EF (1967) Why growth rates differ: postwar experience in nine western countries. Brookings Institution, Washington
- Ellison G, Glaeser E (1999) The geographic concentration of industry: does natural advantage explain agglomeration? *Am Econ Assoc Papers Proc* 89(2):311–316
- Feldman M, Audretsch D (1999) Innovation in cities: science based diversity, specialization and localized competition. *Euro Econ Rev* 43(2):409–429
- Florida R (2005) *Cities and the creative class*. Routledge, New York
- Giesen C, Südekum J (2011) Zipf's law for cities in the regions and the country. *J Econ Geogr* 11(4):667–686
- Glaeser EH, Kallal D, Scheinkman J, Schleifer A (1992) Growth in cities. *Journal of Political Economy* 100(6):1126–1152
- Groot de H, Poot J, Smit M (2009) Agglomeration externalities, innovation and regional growth: theoretical perspectives and meta-analysis. In: Capello R, Nijkamp P (eds) *Handbook of regional growth and development theories*. Edward Elgar, Cheltenham, pp 256–281
- Grossman G, Helpman E (1991a) *Innovation and growth in the global economy*. The MIT Press, Cambridge
- Grossman G, Helpman E (1991b) Endogenous product cycles. *Econ J* 101(409):1214–1229
- Henderson V (2006) Urbanization and growth. handbook of economic growth. In: Aghion P, and Durlauf S (eds) *Handbook of economic growth*. Elsevier-North Holland, New York, pp 1543–1591
- Kaldor N (1963) Capital accumulation and economic growth. In: Lutz A, Hague D (eds) *Proceedings of a conference held by the international economics association*. Macmillan, London, pp 177–222
- Krugman P (1991) Increasing returns and economic geography. *J Polit Econ* 99(3):483–499
- Lucas R (1988) On the mechanics of economic development. *J Monet Econ* 22(1):3–42
- Romer P (1990) Endogenous technological change. *J Polit Econ* 98(5):S71–S102
- Solow R (1957) Technical change and the aggregate production function. *Rev Econ Stat* 39(3):312–330
- Temple J (1999) The new growth evidence. *J Econ Lit* 37(1):112–156

Incorporating Space in the Theory of Endogenous Growth: Contributions from the New Economic Geography

12

Steven Bond-Smith and Philip McCann

Contents

12.1	Introduction	214
12.2	A Simple Model of Endogenous Growth	215
12.2.1	Demand	216
12.2.2	Production	217
12.2.3	Research and Development	218
12.2.4	Equilibrium	219
12.2.5	Balanced Growth	220
12.3	A Two-Region Model of Growth	221
12.3.1	Incorporating Space in the Theory of Growth	222
12.3.2	Model Description	224
12.3.3	Long-Run Location	225
12.4	Spatial Consequences for Economic Growth	228
12.4.1	Integration	228
12.4.2	Growth in Varieties	228
12.4.3	Consumption Growth	229
12.4.4	Agglomeration, Freeness of Trade, and Economic Growth	230
12.4.5	Impact of Knowledge Spillovers upon Economic Growth	231
12.5	Variations to Incorporating Space in the Theory of Growth	231
12.5.1	Mobility of Labor and Capital	231
12.5.2	Vertically Linked Industry	233
12.5.3	Other Characteristics	234
12.6	Conclusions	234
	References	235

S. Bond-Smith (✉)

Department of Economics, University of Waikato, Hamilton, New Zealand
e-mail: sbondsmith@gmail.com

P. McCann

Department of Economic Geography, University of Groningen, Groningen, The Netherlands
e-mail: p.mccann@rug.nl

Abstract

We describe how endogenous growth theory has now incorporated spatial factors. We also derive some of the policy implications of this new theory for growth and economic integration. We start by reviewing the product variety model of endogenous growth and discuss similarities with modeling techniques in the new economic geography. Both use Dixit-Stiglitz competition. Increasing returns provide an incentive for innovation in endogenous growth theory, and in combination with transport costs, increasing returns provide an incentive for firm location decisions in the new economic geography. Since innovation is the engine of growth in endogenous growth models and knowledge spillovers are a key input to innovation production, we also explore how innovation and knowledge have distinctly spatial characteristics. These modeling similarities and the spatial nature of knowledge spillovers have led to space being incorporated into the theory of endogenous growth. We guide the reader through how space is modeled in endogenous growth theory via the new economic geography. Growth by innovation is a force for agglomeration. When space is included, growth is enhanced by agglomeration because of the presence of localized technology spillovers. We consider the many other spatial factors included in models of space and growth. We explore the spatial effects on economic growth demonstrated by these theoretical models. Lastly, we consider policy implications for integration beyond lowering trade costs and discuss how lowering the cost of trading knowledge is a stabilizing force and is growth enhancing.

12.1 Introduction

Theoretical models of endogenous growth explain the engine of economic growth with intentional investments in innovation motivated by monopolistic competition. But these theories have typically ignored space. Endogenous growth and the new economic geography (NEG) have grown quite separately despite similarities in modeling using Dixit and Stiglitz (1977) preferences. Within the literature on innovation, contributions on systems of innovation and the geography of innovation (Audretsch and Feldman 1996) have the potential for a number of spatial aspects to also be incorporated into the theory of growth. More recently, endogenous growth theory has been combined with the NEG and provided insights on how geographic space can influence economic growth.

There are persistent differences in growth rates and incomes between even highly integrated regions such as the European Union or the United States. Endogenous growth theory offers some explanations for varying growth rates. Firms invest in research and development (R&D) to design new innovations, whereby knowledge of existing products is an integral input to R&D. Profits provide an incentive for investment and are protected by patents. The theory implies that varying rates of economic growth may be caused by regions specializing in different sectors with varying rates of productivity or rates of innovation and by

differing institutions that protect patents. The theory fails to provide an adequate explanation of varying growth rates because it does not explain differences in levels of innovation when regions have similar institutions or innovations that are not protected by institutions (e.g., process innovations, firm structure). Spatial factors offer some explanation for these differences, but economic growth theory typically ignores the role of space in determining economic growth outcomes.

Kaldor (1970) explains how trade can drive apart even identical regions as industry agglomerates in a single location. Some contributions to endogenous growth theory include this trade mechanism (Lucas 1988; Grossman and Helpman 1991b, 1995) but still ignore the role of space (distance-related factors) in economic growth. Despite the increasing use of space in economic theory through developments in new trade theory (Krugman 1979), new economic geography (Krugman 1991), and similarities in modeling, it has only been a recent development to incorporate geographic space into growth theory to create spatial models of endogenous growth (Martin and Ottaviano 1999). Developments from the NEG have now led to the incorporation of spatial factors related to both production and knowledge into theoretical growth models. These types of models may help explain varying growth rates between even highly integrated regions with similar institutions. For example, McCann (2009) suggests an economic geography perspective of New Zealand might help explain the difference in growth rates with Australia.

Hence, the new economic geography and growth (NEGG) literature incorporates space into the theory of growth by combining endogenous growth theory with the NEG. This chapter starts by describing the basic theory of endogenous growth (Romer 1990; Grossman and Helpman 1991a) followed by a typical NEGG approach where the theory accounts for the spatial factors of transport costs, migration, and imperfect knowledge spillovers. We review the contribution of these types of spatial models and variations in the use of spatial parameters and discuss the consequences for regional growth policy.

12.2 A Simple Model of Endogenous Growth

Endogenous growth theory uses increasing returns as an incentive for firms to make intentional investments to develop innovations. In all theoretical models of growth, the accumulation of capital (physical and human capital) is the engine of growth. Romer (1990), Lucas (1988), and Aghion and Howitt (1992) treat investment in innovation as investment in an additional type of capital, with increasing returns. While accumulation of physical and human capital suffers from diminishing returns, returns to investment in innovation are not diminishing and growth is sustained in the long run. These theoretical models are separated into two groups: Grossman-Helpman-Romer models (Romer 1990; Grossman and Helpman 1991a) use a love of variety with Dixit and Stiglitz (1977) competition and an increasing number of varieties as the source of growth. Alternatively, Schumpeterian growth models (Aghion and Howitt 1992) use creative destruction or quality ladders where higher-quality products replace existing varieties.

In this section, we present a simple product variety model of endogenous growth through research and development (Romer 1990; Grossman and Helpman 1991a). In subsequent sections, we explore contributions that add space to this model of endogenous growth. To focus on innovation, the models here overlook factor accumulation (such as investments in physical and human capital) so that all investment is in the form of creating new technologies (innovations). In the product variety model, growth comes from an expanding variety of goods. We treat these as final goods using Dixit-Stiglitz preferences as in Grossman and Helpman (1991a). In contrast, Romer (1990) has an expanding variety of intermediate goods which are used to make a final good with a Dixit-Stiglitz production function. Grossman and Helpman (1991a) acknowledge the alternative Dixit-Stiglitz specification of the production function rather than the utility function. The outcomes of the model are essentially the same. We use the final goods version for consistency with the global and local spillover models in Baldwin et al. (2003).

There are two sectors – final goods and the R&D sector. Labor is either employed in producing final goods or in R&D which produces new designs. Workers are free to choose the sector in which they are employed and supply their labor inelastically. Consumers have a taste for diversity and are made better off by an expanding number of varieties. For each new good, there is a sunk cost of innovation that occurs once, when the product is developed. Each firm must first obtain a design from the R&D sector, but once the design is obtained, the firm can produce that variety forever at a constant marginal cost. The presence of fixed costs leads to monopolistically competitive markets. The up-front cost is financed by monopoly profits that are later earned in sales.

12.2.1 Demand

The representative consumer is infinitely lived and has intertemporal preferences:

$$U = \int_{t=0}^{\infty} e^{-\rho t} \ln C_t dt \quad (12.1)$$

where C_t is the consumption index of goods, ρ is the rate of time preference, and time is indexed by t (for simplicity, the subscript t will be dropped hereafter where the time dimension is clear). Consumers have constant elasticity of substitution (CES) preferences over the continuum of final goods $[0, K]$:

$$C = \left[\int_0^K c_i^{\frac{\sigma-1}{\sigma}} di \right]^{\frac{\sigma}{\sigma-1}}, \sigma > 1 \quad (12.2)$$

where c_i is the consumption of variety i and σ is the constant elasticity of substitution. Consumers have a taste for diversity over an infinite set of products $i \in [0, \infty]$ where at any point in time, a subset K is available in the marketplace.

Consumers allocate income between consumption and savings and distribute consumption across available varieties. Intertemporal utility optimization implies that expenditure changes over time according to the Euler equation:

$$\frac{\dot{E}}{E} = r - \rho \quad (12.3)$$

where E is consumer expenditure, \dot{E} is expenditure differentiated with respect to t , and r is the risk-free rate of return on savings. In equilibrium, we have $r = \rho$, $\forall t$. Subject to the budget constraint

$$\int_0^K c_i p_i di \leq E \quad (12.4)$$

consumers allocate expenditure across varieties to maximize utility. With aggregate consumption defined as $C = \left[\int_0^K c_i^{\frac{\sigma-1}{\sigma}} di \right]^{\frac{\sigma}{\sigma-1}}$ in the utility function, we define the price index as $P = \left[\int_0^K p_i^{1-\sigma} di \right]^{\frac{1}{1-\sigma}}$, where p_i is the price of variety i , so that $E = CP$. It can be shown (e.g., Appendix 2.A of Baldwin et al. 2003) that the demand function facing an individual firm is

$$c_i = Ep_i^{-\sigma} P^{\sigma-1} \quad (12.5)$$

Hence, demand is inversely related to relative price.

12.2.2 Production

There are L workers/consumers. Each worker produces one unit of labor per period and supplies its labor inelastically. We assume that each firm takes other firms' prices as given, and with a large number of varieties, firms ignore the effect of their own price on the price index. Some portion of labor $L_M < L$ is employed in the manufacturing sector. Remaining labor is employed in the innovation sector L_I such that $L_I + L_M = L$. Firms choose optimal prices in order to maximize profit $\pi_i = (p_i - \beta w) c_i$, where β describes the marginal units of labor per unit of production and w describes the wage, such that βw is the marginal cost of production. This implies that optimal prices are a constant markup over marginal cost:

$$p_i = \frac{\sigma}{\sigma - 1} \beta w, \quad \forall i \in [0, K] \quad (12.6)$$

It is possible to impose a normalization $\beta = \frac{\sigma-1}{\sigma}$, which implies that $p_i = w$. We avoid this normalization here so that the reader can identify in the formulas how the productivity of labor in both manufacturing and R&D is important for the allocation of labor between sectors. As there are symmetric demands for all varieties, every

firm in the manufacturing sector yields the same price, production, and operating profit per period. We normalize expenditure $E = 1 \forall t$. With $E = 1$ and all firms being otherwise identical, individual firm profit in each period is given by

$$\pi_i = \frac{E}{K\sigma} = \frac{1}{K\sigma} \quad (12.7)$$

12.2.3 Research and Development

A manufacturing firm has a one-off fixed cost to develop the patent to the good (or purchase it from an entrepreneur) in the R&D sector which generates designs for new varieties of final goods. Each variety requires one unit of knowledge capital produced by the R&D sector. Individual firms face an innovation cost of a_I units of labor per unit of capital produced. We follow Grossman and Helpman (1991a) and Romer (1990) using a learning curve such that the marginal cost of new knowledge capital a_I declines as cumulative knowledge output increases. Romer (1990) rationalizes this assumption by referring to the non-rival nature of knowledge, emphasizing the role of knowledge spillovers. Labor and the stock of knowledge (equal to the number of varieties) are used to develop new innovations. Each innovation adds to the stock of knowledge that can be used for developing future innovations. Innovation production is given by

$$\dot{K} = \frac{L_I}{a_I}, \quad F = wa_I, \quad a_I = \frac{1}{K_t} \quad (12.8)$$

where \dot{K} is knowledge capital differentiated over time, L_I is the labor employed in the innovation sector, F is the fixed cost of innovation to develop a new variety in the R&D sector, and $a_I = \frac{1}{K_t}$ describes the productivity of the R&D sector increasing with cumulative output (i.e., the fixed cost of each innovation decreases over time). The model is based on Grossman and Helpman (1991a), but for consistency, the functional form used here is adapted from Baldwin and Forslid (2000). From Eq. (12.8), it follows that the rate of growth in the number of varieties is equal to L_I , which may be scaled by a constant to calibrate the model as described in Grossman and Helpman (1991a).

An entrepreneur seeks funding for up-front costs (R&D wages) from credit markets (or provides that credit in foregone wages). We assume there are no frictions in credit markets and no aggregate uncertainty so the purchasing of a patent can also be thought of as the entrepreneur issuing debt or equity (or some combination). Once a patent has been obtained, the manufacturer has monopoly rights to produce variety i forever at constant marginal cost. Equity owners are paid the infinite stream of profits from the firm. Free entry into the research sector implies that labor is hired such that its wage equals its marginal product. At time t and with constant interest rates, the present value of the future stream of profits, v_t , is

$$v_t = \int_{s=t}^{\infty} \pi_s e^{-r(s-t)} ds \quad (12.9)$$

Differentiating and rearranging, we find the no arbitrage condition:

$$\dot{v}_t = -e^{-r(s-t)} \pi_s + rv_t \quad (12.10)$$

This can also be written as a rate of return, $\frac{\dot{v}_t}{v_t} + \frac{\pi_t}{v_t} = r$. The “no arbitrage” condition describes that in the interval between t and $t + dt$, the owners of the patent (equity holders) receive a return (made up of the profit rate $\frac{\pi_t}{v_t}$ and the rate of capital gain (loss) $\frac{\dot{v}_t}{v_t}$) equal to the yield on a riskless loan. In other words, for a manufacturing firm to purchase a patent (or investors to hold equity/debt), the payoff must exceed the opportunity cost.

The cost of research that yields $\dot{K} = \frac{l_t}{a_I}$ incremental varieties is wa_I and has the value $v_t \dot{K} = v_t \frac{l_t}{a_I}$, where l_I is the labor input by a typical entrepreneur. Continuous growth, $\dot{K} > 0$, involves an active research sector, and free entry requires the research costs to be equal to the value of research for all t . If the costs of research are greater than the value of R&D, no research would occur in equilibrium. A situation where the cost of research is less than the value of R&D will never occur in equilibrium because it would cause an unbounded demand for research labor. Equilibrium therefore requires $v_t \leq wa_I$ with equality when $\dot{K} > 0$.

12.2.4 Equilibrium

Rather than deriving equilibrium, we just describe the equilibrium or steady state. For a full discussion of equilibrium conditions, see Grossman and Helpman (1991a), Baldwin and Forslid (2000), or Baldwin et al. (2003). In equilibrium, we have a flow of new innovations:

$$\dot{K} = \begin{cases} \frac{L}{a_I} - \frac{\beta}{\bar{v}} & \text{for } v > \bar{v} = \frac{\sigma - 1}{\sigma} \frac{a_I}{L} \\ 0 & \text{for } v \leq \bar{v} = \frac{\sigma - 1}{\sigma} \frac{a_I}{L} \end{cases} \quad (12.11)$$

with $L = L_I + L_M$, i.e., total employment is the sum of R&D employment and manufacturing employment. Substituting the interest rate $r = \rho$ and the profit rate $\pi = \frac{1}{K\sigma}$ into the no arbitrage condition, the change in firm value is a function of the value of a firm and the number of firms:

$$\dot{v} = \rho v - \frac{1}{K\sigma} \quad (12.12)$$

These two differential equations, Eqs. (12.11) and (12.12), describe the dynamic equilibria.

12.2.5 Balanced Growth

If conditions allow for employment in R&D, there are an increasing number of varieties. As firms compete for a fixed supply of labor, the output per firm and the value of a firm go down over time. Research into new varieties remains profitable since the cost of innovation decreases as the number of varieties increases. We denote the steady growth rate of the number of varieties, $\frac{\dot{K}}{K}$, by g_K . If we define a new variable $V = \frac{1}{K^v}$, representing the inverse of the economy's aggregate equity value, the growth rate is

$$g_K = \frac{\dot{K}}{K} = \begin{cases} L - \beta V & \text{for } V < \frac{\sigma}{\sigma-1} \frac{L}{a_I} \\ 0 & \text{for } V \geq \frac{\sigma}{\sigma-1} \frac{L}{a_I} \end{cases} \quad (12.13)$$

These definitions also imply $\frac{\dot{V}}{V} = -g_K - \frac{\dot{v}}{v}$. By substitution of $\dot{v} = \rho v - \frac{1}{K^\sigma}$, we find

$$\frac{\dot{V}}{V} = \frac{1}{\sigma} V - g_K - \rho \quad (12.14)$$

The model is reduced to one differential equation, and the condition for growth is given by Eq. (12.13). We can calculate the steady state rate of innovation by setting $\dot{V} = 0$:

$$g_K = \frac{L}{\sigma} - \beta \rho \quad (12.15)$$

This is positive, so long as $L > \rho(\sigma - 1)$; otherwise, growth is zero. Growth is positively related to the scale of the economy (L), which is a common property of these models. Innovation (and incentives for R&D investment) is sustained because there are offsetting forces of declining profits due to expanding varieties and falling product development costs due to research externalities.

This is not the overall growth rate of the economy. To understand macroeconomic growth, we are interested in the growth rate of the consumption index, $C = \left[\int_0^K c_i^{\frac{\sigma-1}{\sigma}} di \right]^{\frac{\sigma}{\sigma-1}}$. Since $E = CP = 1 \forall t$, growth is also the rate at which the price index $P = \left[\int_0^K p_i^{1-\sigma} di \right]^{\frac{1}{1-\sigma}}$ declines. The growth rate of consumption g_C can be shown to be $g_C = \frac{g_K}{\sigma-1}$. This is also not GDP growth. GDP is defined as the value added in both manufacturing and R&D. GDP grows at a rate equal to a weighted average of the growth rates of the index of manufacturing output/consumption and of research output. Since R&D is usually only a small percentage of a country's GDP, the difference is negligible. See Grossman and Helpman (1991a, p. 63) for a discussion.

12.3 A Two-Region Model of Growth

Virtually all endogenous growth models rely on technical externalities such as knowledge spillovers and production externalities. Endogenous growth models usually assume a frictionless spillover of knowledge. The reality is that knowledge is not transferred so effortlessly. While some knowledge can be codified and transferred easily, much knowledge is at least partially tacit. Spillovers of tacit knowledge occur over space and time through face-to-face contact (McCann 2007) and migration (Faggian and McCann 2009). Eaton and Kortum (1999) show that knowledge spillover and production externalities are related to the geographic distribution of manufacturing and R&D. A better understanding of the economics of innovation (Nelson 1993) and its geographic characteristics (Audretsch and Feldman 1996) significantly improves our understanding of economic growth.

Innovation is a predominantly local event and is now included in economic geography. Acs and Varga (2002) note the similarities between modeling techniques of endogenous growth theory and the new economic geography, suggesting a new model of technology-led regional economic development that combines the two fields with insights from the economics of innovation. Knowledge and innovation also have space, time, and cost characteristics in their spillover between locations. This role of space and time in knowledge spillovers means economic growth also has spatial characteristics.

Given this understanding of innovation, the concentration of economic activity also results in greater knowledge spillovers between firms in concentrated locations. In endogenous growth literature, there is recognition of partial international knowledge spillovers. Grossman and Helpman (1991b) model foreign knowledge as an innovation input in a small economy where the availability of foreign knowledge is dependent on the level of trade, yet these models ignore the role of space. Space can be added to the theory of growth by including spatial characteristics in knowledge spillover inputs to innovation production.

Transport costs are also a key spatial parameter typically ignored in endogenous growth models. The new trade theory (Krugman 1979) and the new economic geography (NEG) (Krugman 1991) include transport costs and have Dixit-Stiglitz competition in common with many theoretical endogenous growth models. Transport costs can therefore be included easily within endogenous growth. The result of transport costs is the concentration of production in specific locations, when transport costs reach some low threshold. This is known as the core-periphery model. With low enough transport costs, firms choose to locate close to their customers to reduce transport costs. When models also allow for migration, workers choose to locate near producers to reduce their cost of living. These transport cost-related phenomena are known as the home market effect because it causes the concentration of firms and people.

Higher transport costs may induce firms to seek locations where there are fewer firms to compete with. This is known as the market crowding effect. It is the balance of these two effects that determines equilibrium and the steady state. Concentration

occurs at low transport costs when the home market and cost of living effects dominate the market crowding effect, while dispersion occurs at higher transport costs, where market crowding dominates. The NEG suggests that imperfect integration may create regional winners and losers (Krugman 1991; Krugman and Venables 1995). A particularly interesting characteristic is that the economic conditions of two regions can be exactly the same yet yield dramatically different economic outcomes.

12.3.1 Incorporating Space in the Theory of Growth

New economic geography and growth (NEGG) models combine horizontal innovations à la Grossman-Helpman-Romer with the NEG (e.g., Baldwin et al. 2001; Baldwin and Forslid 2000; Martin and Ottaviano 1999; Fujita and Thisse 2003) predominantly due to the fundamental use of Dixit-Stiglitz competition. Different NEGG models vary assumptions on the mobility of capital, labor, and industry or consumer demand to influence the forward and backward linkages. Here, we describe a typical NEGG modeling approach (Baldwin and Forslid 2000) that includes the spatial factors:

- Location
- Migration
- Transport costs
- Local knowledge spillovers
- Imperfect global knowledge spillovers

The model has two regions that trade. There is a traditional goods sector with perfect competition that employs immobile unskilled workers L_T . Consumers have a taste for traditional goods such that $C = C_M^\mu C_T^{1-\mu}$, where C_M is the index of manufactured goods (similar to C in the previous section) and C_T is the traditional goods sector. Foreign region variables are denoted by an asterisk (*). The representative consumer is infinitely lived and has intertemporal preferences:

$$U = \int_{t=0}^{\infty} e^{-\rho t} \ln [C_{Mt}^\mu C_{Tt}^{1-\mu}] dt \quad (12.16)$$

In what follows, the time subscripts will again be suppressed for simplicity. Transport costs are zero in the traditional goods sector, and workers in this sector cannot migrate between regions. In the real world, workers in the traditional goods sector are not necessarily unskilled or immobile. The important feature here is that the factor of production for traditional goods is immobile, and “unskilled” is the commonly used term in these models. The purpose of the additional sector in this model is that some residual demand remains in the periphery, even when there is full agglomeration, so that regions continue to trade.

Skilled workers (L_K) are employed in either manufacturing or innovation (similar to workers in the previous section with subscript K since they work in the knowledge sectors of manufacturing or innovation). The world population of

skilled and unskilled workers is normalized to one such that $L = L_K + L_T = 1$. Skilled workers and manufacturing firms have a choice of location. Skilled workers respond to wage pressure when making a decision to migrate between regions. If there are differences in real wages, there will be migration. The perfect price index describes the price index of utility and therefore includes traditional goods such that $P \equiv P_T^{1-\mu} P_M^{\frac{\mu}{\sigma-1}}$. The change in skilled workers in the home region is given by the ad hoc migration equation in Fujita et al. (1999):

$$\dot{L}_K = (\omega_K - \omega_K^*) s_H (1 - s_H) \quad (12.17)$$

$$s_H = \frac{L_K}{L_K + L_K^*}, \quad \omega_K = \frac{w}{P}, \quad \omega_K^* = \frac{w^*}{P^*} \quad (12.18)$$

where \dot{L}_K is skilled labor in the home region differentiated over time, s_H is the share of skilled workers in the home region, and ω_K is the real wage of skilled workers in the home region. Since the real wage is defined by means of the perfect price index, workers migrate to the region that provides the highest level of utility.

Manufactured goods transported between regions incur transport costs that take Samuelson's "iceberg" form where transport costs are incurred in the good itself. The manufacturer produces more of the good than actually arrives because some portion of the good "melts" in transit. If τ represents the proportion of the final good that arrives at the destination, the remaining portion is used up during transportation. Hence, $\tau < 1$ is a measure of the freeness of trade or an index of the inverse of transport costs. Transport costs for the traditional goods sector are assumed zero ($\tau = 1$). Firms are incentivized to locate in the largest market to minimize transport costs. From the migration equation above, skilled workers try to locate in the region with more firms as this reduces their cost of living (since they have a taste for diversity) by increasing real wages.

So far, we have added space with migration and transport costs which affect manufacturing, but we now also add space to innovation production. Since knowledge does not transfer completely between regions, not all knowledge is available to entrepreneurs when manufacturing is shared between regions. Innovation is included in the manufacturing sector the same as in the endogenous growth model of Sect. 12.2 but now with partial spillovers of knowledge between regions. Individual firms face the innovation cost of a_I units of labor for each unit of knowledge capital produced. Innovation production in the home region is given by

$$\dot{K} = \frac{L_I}{a_I}, \quad F = w a_I, \quad a_I = \frac{1}{K + \lambda K^*}, \quad 0 \leq \lambda \leq 1 \quad (12.19)$$

where \dot{K} is knowledge capital differentiated over time, L_I is the skilled labor employed in the innovation sector, λ is the ability for foreign knowledge to be used in the home region, and $a_I = \frac{1}{K + \lambda K^*}$ describes how productivity of the R&D sector increases with cumulative output. Hence, the model assumes perfect local knowledge spillovers but imperfect spillovers between regions. The parameter λ

represents how space affects knowledge production such that firms choose a location that considers how existing knowledge can be used for innovation. In this way, firms are attracted to regions where other firms are located because the cost of innovation is lower.

12.3.2 Model Description

Consider the product variety model of the previous section together with these additional spatial factors. Again, we normalize world expenditure $E_w = 1, \forall t$. Subject to the budget constraint, consumers allocate expenditure across varieties to maximize utility. Hence, in the home region, $P_M C_M + P_T C_T \leq E$, where P_M is the local price index of manufactured goods (the world equivalent is a weighted average price index such that $P_M C_M + P_T C_T \leq E_w$) and P_T is the price of traditional goods. Consumers spend a constant portion of their expenditure on manufactured goods and the rest on traditional goods:

$$P_M C_M = \mu E, \quad P_T C_T = (1 - \mu)E \quad (12.20)$$

Total expenditure on traditional goods is equal to $(1 - \mu)E_w$. The traditional goods sector is perfectly competitive, with 1:1 technology (one unit of unskilled labor input yields one unit of traditional goods output) and constant returns to scale. Total production of traditional goods is shared across both regions. Let L_T and L_T^* be the supply of unskilled workers in the home and foreign regions, respectively. We follow Krugman (1991) and set the worldwide stock of skilled workers to μ and the stock of unskilled workers to $(1 - \mu)$ shared equally between regions:

$$L_T = \frac{1 - \mu}{2}, \quad L_T^* = \frac{1 - \mu}{2} \quad (12.21)$$

The choice of units ($1 - \mu$ unskilled workers and μ skilled workers) follows Krugman (1991) and ensures that prices and wages in the traditional goods sector are the numéraire and that the nominal wage rate of skilled workers equals that of unskilled workers. If the number of skilled workers was specified differently, the wages of skilled workers are a constant multiple of the wage rate of unskilled workers. We maintain simplicity by avoiding this additional multiple. A scaling factor could also be used to calibrate the model to any arbitrary growth or wage rate.

Unskilled workers provide one unit of production per period, i.e., $\int_0^{L_T} C_T + \int_0^{L_T^*} C_T^* = (1 - \mu)$. Free trade ensures the same nominal price of traditional goods and equal nominal wages in the two regions. With full employment of $1 - \mu$ unskilled workers and 1:1 technology, the traditional goods sector is the numéraire:

$$\begin{aligned} w_T \left(\int_0^{L_T} C_T + \int_0^{L_T^*} C_T^* \right) &= w_T(1 - \mu) = P_T \left(\int_0^{L_T} C_T + \int_0^{L_T^*} C_T^* \right) \\ &= P_T(1 - \mu) = (1 - \mu)E_w \end{aligned} \quad (12.22)$$

$$w_T = P_T = w_T^* = P_T^* = 1 \quad (12.23)$$

The remainder of the analysis focuses on the manufacturing sector. The home region produces K manufactured varieties and the foreign region produces K^* varieties. Consumers have a CES preferences over the continuum of manufactured goods $[0, K + K^*]$, such that

$$C_M = \left[\int_{i=0}^{K+K^*} c_i^{\frac{\sigma-1}{\sigma}} di \right]^{\frac{\sigma}{\sigma-1}}, \quad \sigma > 1 \quad (12.24)$$

where c_i is the consumption of variety i and σ is the constant elasticity of substitution. Defining the local price index of manufactured goods as in the model of Sect. 12.2, $P_M = \left[\int_0^{K+K^*} p_i^{1-\sigma} di \right]^{\frac{1}{1-\sigma}}$ where p_i is the price paid by local consumers, the demand function in the home region facing an individual manufacturer is $c_j = \mu E p_j^{-\sigma} P_M^{\sigma-1}$, and the equivalent demand function exists in the foreign region with the foreign region's price index.

Manufacturing firms in each region face the same optimization problem as in the endogenous growth model: $\max_{p_i} \pi_i = [p_i - \beta w] c_i$, where βw is the marginal cost of production. Firms ignore the effect of their own price on the index. Once again, optimal prices are a constant markup over marginal cost, and transport costs are passed on directly to consumers:

$$p_i = \frac{\sigma}{\sigma - 1} \beta w, \quad p_i^* = \frac{\sigma}{\sigma - 1} \frac{\beta w}{\tau} = \frac{p_i}{\tau}, \quad \forall i \in [0, K] \quad (12.25)$$

where p_i and p_i^* are the local and export prices of a home manufacturer. A foreign manufacturer has analogous prices, with transport costs on goods exported to the home region. Here, it is also possible to impose the same normalization $\beta = \frac{\sigma-1}{\sigma}$ such that $p_i = w$ and $p_i^* = \frac{w}{\tau} = \frac{\beta w}{\tau}$. While its distribution is subject to worker migration, by following Krugman's (1991) choice of units where the worldwide stock of skilled workers is μ , nominal skilled wages in equilibrium are $w = 1$ or $w^* = 1$ for the core-periphery outcome and $w = w^* = 1$ in the equal distribution outcome.

12.3.3 Long-Run Location

We characterize the long run as a “steady state”: defined by an unchanging growth rate in the number of manufactured varieties, its regional division, as well as the prices and quantities defined by short-run equilibrium above. Migration of

knowledge workers due to spatial inequality of real wages leads to the long-run equilibrium. With the migration equation above and particularly the role of the perfect price index in this equation, we can intuitively see that real wages will only be unequal when one region has a larger share of manufacturing. When this occurs, the larger region is also the lowest cost location for innovation to occur because of greater knowledge spillovers. Furthermore, at low levels of transport costs, there are higher profits in the larger region. At high levels of transport costs and only a slightly unequal equilibrium, there may be higher margins in the smaller region due to the market crowding effect which would return the system to the equal distribution outcome. Through intuition, we can see that there are two long-run types of steady states:

- The equal distribution outcome
- The core-periphery outcome

See Baldwin and Forslid (2000) for a more formal discussion of the conditions of the steady state in the NEGG model here and Baldwin et al. (2003) for a discussion of other NEGG models.

The equal distribution outcome is where both regions have half the skilled workers, half the manufacturing, and half the traditional goods production. The other steady state is the core-periphery outcome where all manufacturing concentrates in a single region (either home or foreign) known as the core and only unskilled workers (the traditional goods sector) remain in the other region known as the periphery. Traditional goods production is split equally between regions.

If there are asymmetric transport costs, it is not inevitable that the region with the lowest transport costs will be the core. The core region will be the one which has the higher share of varieties and where the difference in the number of varieties is large enough to trigger a switch from the equal distribution outcome to the core-periphery outcome. This could be for several reasons. Since every variety has a patent forever, hysteresis plays a large role in determining which region is the core. For example, an initial higher endowment of resources might lead to a greater number of manufacturers and innovators, or greater infrastructure investment at some stage (and temporarily freer trade) might also trigger agglomeration. Similarly, temporarily different policy settings between regions where one region has favorable policies for R&D could lead to initially higher rates of innovation, a greater share of varieties, and agglomeration. While not included in typical NEGG models, stochastic effects could mean one region gets “lucky.” In the model here, innovations are simply costs where each firm has to employ a certain amount of skilled labor in R&D in order to achieve an innovation. In reality, successful innovations are not so guaranteed. The inclusion of probabilistic outcomes in the R&D sector could mean one region achieves a higher rate of innovation by luck, resulting in it becoming the core.

[Figure 12.1](#), reproduced from Baldwin and Forslid (2000) but with a different measure of trade freeness, describes the possible equilibria with different combinations of trade freeness and knowledge diffusion. As the level of trade freeness increases (i.e., transport costs decline), the break point τ_B describes the level of trade freeness where the equal distribution outcome is no longer a steady state. The sustain point τ_S describes the level of trade freeness at which the distribution of

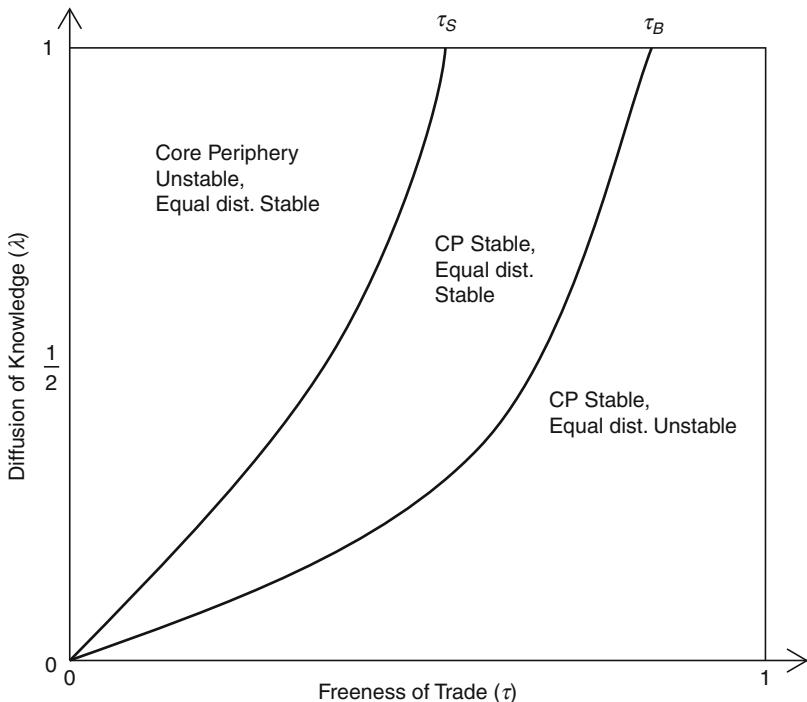


Fig. 12.1 Core-periphery and symmetric equilibrium stability map

firms and workers switches from the core-periphery outcome to the equal distribution outcome when trade freeness is declining (transport costs increase). The values of trade freeness between the sustain and break points represent situations which both the potential equilibria outcomes are stable. As the level of knowledge spillovers λ varies, so do both the break and sustain points. Figure 12.1 describes how the break and sustain points increase as knowledge spillovers increase. Alternatively, Fig. 12.1 describes the combinations of knowledge spillovers and trade freeness that result in stable (and unstable) equilibria for both the equal distribution outcome and the core-periphery outcome. There are three sections within the knowledge spillover (λ) and trade freeness (τ) space. In the top-left corner, the core-periphery equilibrium is unstable and the equal distribution is stable. In this situation, trade freeness is sufficiently low (high transport costs) that the market crowding effect means firms make a higher margin by locating away from other firms. There is very little trade (if any) between regions. Closer to the curve, regions will trade, but the market crowding effect always dominates the home market, cost of living, and innovation cost effects. In the middle section, both the equal distribution and core-periphery equilibrium are stable. If there is an equal distribution, regions will trade, but it is possible that with an external shock, the home market, cost of living, and innovation cost effects could dominate the market crowding effect and the system would switch to the core-periphery outcome. Similarly, if there is a core-periphery

equilibrium, an external shock to the distribution could lead to the market crowding effect dominating the home market, cost of living, and innovation cost effects causing a switch to the equal distribution outcome. Lastly, the bottom right section describes combinations of knowledge spillovers and trade freeness where only the core-periphery outcome is stable. In this situation, the home market, cost of living, and innovation cost effects always dominate the market crowding effect.

12.4 Spatial Consequences for Economic Growth

The incorporation of space in the theory of growth means the model recognizes the role of space through transport costs and through knowledge not transferring perfectly between locations. Let us consider world and regional growth in both the short and long run in the two possible types of equilibria: core-periphery and equal distribution. Because regions are able to trade, even a periphery region benefits from growth in the number of varieties produced in the core. Over time, the price index for manufactured varieties falls as more varieties are invented, and producers of traditional goods experience growth in real income because they trade for manufactured goods. We consider growth in terms of the number of manufactured varieties and growth in terms of the consumption bundles available to all consumers.

12.4.1 Integration

While traditional conceptions of integration refer to lowering of the cost of trading goods, Fig. 12.1 shows that incorporating space and growth gives a more detailed view of integration where we can also view integration as lowering the cost of trading ideas. Integration policies which focus solely on free trade may be destabilizing and result in a deindustrialization of the periphery region. That is, when we lower trade costs alone, the region that emerges as the periphery suffers relative to the region that emerges as the core. Integration policies that also focus on knowledge spillovers (or entirely on knowledge spillovers) will be growth enhancing for both regions. The model shows how this form of integration is stabilizing, while pure trade cost integration can be destabilizing.

12.4.2 Growth in Varieties

The number of manufactured varieties worldwide evolves according to

$$\dot{K} + \dot{K}^* = \frac{L_I}{a_I} + \frac{L_I^*}{a_I^*}, \quad a_I = \frac{1}{K + \lambda K^*}, \quad a_I^* = \frac{1}{\lambda K + K^*} \quad (12.26)$$

For simplicity, we drop the subscript i for p_i because home firms are symmetric and prices are equal for all home firms. Once a blueprint or variety is invented,

manufacturers require β marginal units of labor per unit of production, so aggregate demand for labor in the manufacturing sector in the home region is $\frac{\beta}{p}$. As in the endogenous growth model without space, equilibrium in the skilled labor market in the home region requires $L_K = L_I + L_M = a_I \dot{K} + \frac{\beta}{p}$. In the equal distribution outcome, prices are higher than the core-periphery outcome because of the additional cost to transport goods between regions. A larger share of skilled labor is used in manufacturing because each producer has to produce a larger amount to cover the cost of transport. In other words, the cost of transport increases the marginal cost of production such that some labor is no longer available for innovation. When freeness of trade is greater, i.e., the cost of transport is lower, more labor is available for growth. As such, incorporating space in the theory of growth shows how trade liberalization and agglomeration are growth enhancing for world growth.

Turning to regional growth, the number of manufactured varieties in the home region evolves according to

$$\dot{K} = \frac{L_I}{a_I}, \quad a_I = \frac{1}{K + \lambda K^*} \quad (12.27)$$

Trade liberalization and agglomeration (in the home region) are growth enhancing because they reduce the cost of transport. However, if transport costs induce the core-periphery outcome, there is no manufacturing in the periphery and therefore no growth in varieties produced by that region. That is, reducing transport costs means growth in varieties may be limited to a specific region(s). Therefore, trade liberalization is not growth enhancing for growth in varieties for the region that emerges as the periphery.

For both world and regional growth, the inclusion of space means firms face an innovation cost that is dependent upon location. The output of skilled workers in the innovation sector is greater when knowledge is more available. With s being the home region's share of manufacturing, the rate of growth is $\frac{L_I w}{K+K^*} [s(K + \lambda K^*) + (1 - s)(\lambda K + K^*)]$. That is, when λ is greater, both world and regional growth increase. Including space in the theory of growth shows how closer economic integration is growth enhancing for world and regional growth in varieties. Similarly, when one region has a greater share of manufacturing than the other region, growth increases for the agglomerated region. Agglomeration in either region is growth enhancing for world growth and for regional growth in the region where agglomeration occurs. However, in the core-periphery outcome, there is zero growth in the number of varieties in the periphery region as no varieties are manufactured there, no skilled workers are employed, and no innovation occurs.

12.4.3 Consumption Growth

While so far we have described the effect of space on the growth rate of the number of varieties, this is not the overall growth rate because we have ignored traditional goods. In considering the growth rate of the overall economy, we are interested in

the growth rate of what people actually consume. In other words, we are interested in what the income to workers allows those consumers in each region to purchase which is measured by the growth rate of the consumption index, C , where $E = CP = 1$. This best describes how the well-being of consumers increases over time. While there is no growth in the number of varieties produced in the periphery, the ability to trade traditional goods for manufactured goods allows the unskilled workers to benefit from innovations in the core.

Since $E = CP$, the rate at which the consumption index grows is the rate at which the perfect price index declines. In the endogenous growth model of Sect. 12.2, the growth rate of consumption g_c was shown to be $g_c = \frac{g_k}{\sigma-1}$. With the addition of the traditional goods sector, the overall perfect price index is to a power of $\frac{(\sigma-1)}{\mu}$. The perfect price index is falling at a rate of $g_c = \frac{ug_k}{(\sigma-1)}$. Notably, the growth rate of consumption is the same in both regions whether we have a symmetric outcome or the manufacturing concentration outcome. This is because the price index for both regions falls at the same rate, since consumers in both regions still spend the same portion of their earnings on traditional goods – in the steady state, the growth rate of consumption is equal in both regions. The inclusion of space does not explain the differences in growth rates between locations in the long run. Instead, space affects the world rate of growth and the share of wealth/earnings in each location.

In the short run, however, there can be different growth rates between locations if the regions are in transition between steady states. Given $\tau < 1$, the price index will be permanently lower in a core location because core location consumers do not pay transport costs for manufactured goods. If the economies are shifting from an equal distribution to the core-periphery outcome, growth rates in the periphery will be temporarily lower (or even negative) as periphery consumers transition to paying transport costs on a greater share of the manufactured goods they consume (eventually all goods). Consumers in the core gradually pay transport costs on a smaller share of manufactured goods, and the core will have higher growth rates.

12.4.4 Agglomeration, Freeness of Trade, and Economic Growth

Agglomeration is growth enhancing in the long run through both transport costs and knowledge spillovers. Agglomeration minimizes the total cost of transport if all manufacturing and the majority of consumption is in one location. Agglomeration is also growth enhancing because it increases knowledge spillovers if all R&D occurs in one location.

Increased freeness of trade is growth enhancing in the long run, but in the short run, the outcome is ambiguous. Increased freeness of trade is always growth enhancing if there is no change in the distribution of economic activity. However, as described in Fig. 12.1, increased freeness of trade can lead to a switch from the equal distribution outcome to the core-periphery outcome. While this is significantly growth enhancing for the region that becomes the core, it is temporarily growth diminishing for the periphery, while the two regions transition to the new equilibrium.

12.4.5 Impact of Knowledge Spillovers upon Economic Growth

Knowledge spillovers are generally growth enhancing. Increased knowledge spillovers mean firms have a lower cost of innovating, and therefore, there is a greater growth rate in varieties and consumption. If we are in the core-periphery equilibrium, increasing knowledge spillovers has no effect on growth because knowledge is unaffected by space since all production is in a single location.

However, as with agglomeration, the effect is ambiguous if there is a change in the steady state. A large enough increase in knowledge spillovers could lead to a switch from the core-periphery outcome to the equal distribution outcome (see Fig. 12.1). With a change in the location of production from one region to multiple regions, the knowledge spillover parameter now has an effect on growth when there was previously no effect. That is, firms initially had access to all knowledge because all manufacturing was in the same region, but in the new steady state, foreign knowledge is only partially available. While knowledge spillovers are generally growth enhancing, there is the possibility of knowledge spillovers being growth reducing in the former core region if it brings about the sharing of manufacturing.

If we consider the steady state where production is shared between locations, knowledge spillovers are growth enhancing. Furthermore, knowledge spillovers also make production in the equal distribution outcome more stable. That is, increasing knowledge spillovers means changes in trade costs are less likely to lead to a switch to the core-periphery outcome (see Fig. 12.1). With greater knowledge spillovers, production in both regions is a stable equilibrium for a greater range of trade freeness.

12.5 Variations to Incorporating Space in the Theory of Growth

In the NEGG literature, there are many variations of the model presented here. These include differences in the mobility of labor or capital, the inclusion of intermediate goods, heterogeneous firms, multiple labor types, and heterogeneous skill levels. Other areas of economics also incorporate space by using continuous space (rather than discrete regions), by defining location on an interval, by incorporating land as a factor of production, and by introducing congestion costs. All of these variations have different effects on the role of space, location, and geography on growth, but in general, incorporating space in the theory of growth has similar effects to those presented here.

12.5.1 Mobility of Labor and Capital

The model here describes the typical approach by NEGG scholars to incorporate space in the theory of growth with the inclusion of migration of skilled labor. The effect of footloose skilled labor can lead to catastrophic agglomeration, which

means the model is unable to show other unequal internal steady states. We describe the model that includes skilled worker migration to demonstrate the role of firm and worker location choices and how migration influences innovation. Highly skilled workers and innovators are internationally mobile, so it is important to consider how this affects the location of innovation and subsequently economic growth.

Capital mobility is the ability for capital to shift between locations. In all endogenous growth models, growth comes from the accumulation of capital. Capital can come in a number of forms: human capital, physical capital, or knowledge capital. We think of labor and education as human capital, which is able to migrate between locations in the model above. Physical capital is the equipment used in production such as machinery and production plants. This has been excluded from the model above. Knowledge capital is the ideas generated in the innovation sector which are marketable and tradable through patents. This is the type of capital commonly modeled in endogenous growth and NEGG literature.

There are two options for the mobility of knowledge capital. With mobile capital, the owners of capital can decide where to locate production. If knowledge capital is mobile, the number of innovations produced (and owned) by one region may be different from the number of firms actually producing in that region. That is, the developer of a patent may choose to produce in a region other than their own. In this situation, the decision to accumulate capital is the same in all locations; the mobility of capital eliminates demand-linked causality such that the shifting of production does not shift the location of consumption or the earnings from owning a manufacturing firm. Alternatively with immobile capital, the owners of capital are only able to produce within the region where they are located. With immobile capital, any shift that favors production in one location leads to new capital in that region. Since owners are local, this also leads to expenditure shifting and further production shifting via the home market effect.

In many NEGG models, such as in Martin and Ottaviano (1999), Baldwin et al. (2001), and Baldwin and Martin (2004), migration is not allowed. In these models, workers are instead completely mobile between traditional manufacturing and innovation sectors but not between regions. These models require an extra assumption that a single country's labor endowment must not be enough to meet global demand for traditional goods, to avoid complete specialization in manufacturing goods only.

In models with labor immobility and capital mobility, when we reach the steady state, the owners of capital are indifferent between producing in either region. With localized knowledge spillovers, however, innovators prefer to be located in the region with the highest level of manufacturing. Despite the differences, these models reach similar steady states to the model presented above. In particular, space has the same effects on growth because space is included using the same mechanisms with localized knowledge spillovers and transport costs. Agglomeration is growth enhancing due to localized knowledge spillovers, and knowledge spillovers are growth enhancing because they reduce the cost of innovation.

In models without labor or capital mobility, agglomeration is enabled by either vertical linkages in production or the spatial influence on knowledge creation and

transfer. If NEGG models have immobile capital and mobile labor, these models have the same catastrophic agglomeration described by the model above (and most NEG models) because innovation occurs at a faster rate in a region with greater capital, and this is self-reinforcing as all new firms prefer to innovate in the location with the largest share of manufacturing. Whenever labor is mobile, agglomeration is catastrophic.

However, models with immobile capital and immobile labor offer an alternative advantage of unequal internal solutions. That is, a range of transport costs and knowledge spillovers that yields steady states where one region has a larger share of manufacturing (but not all) than the other. As there is no migration, this means the region with the larger share of manufacturing has a share of traditional goods production smaller than the other region's share of traditional goods production. Even though these models ignore the role of migration in economic activity and growth, it does allow us to consider the effect on knowledge spillovers and growth when there are unequal levels of agglomeration. The effect is very similar to the core-periphery outcome. Growth rates are equal in both regions because consumers in the low manufacturing region still benefit from innovations made in the high manufacturing region because of trade. Similarly, the growth rate in varieties is greater in the high manufacturing region because of localized knowledge spillovers. Real wages are also higher in a high manufacturing region. Without migration, there is no mechanism to equalize real wages between regions.

Another advantage of modeling with labor immobility between regions is there is no need for the modeling trick of the Krugman's (1991) core-periphery model which fixes the share of skilled and unskilled workers. Instead, labor mobility between sectors equalizes real wages between manufacturing and traditional sectors within region, and zero transport costs in the traditional goods sector equalize nominal wages. Even though these alternative models have some features that are mathematically elegant, we chose to explore the model including migration because we view it as a more realistic description of spatial endogenous growth.

12.5.2 Vertically Linked Industry

Other types of NEGG models have vertically linked industry following the practice of some NEG models (Krugman and Venables 1995; Venables 1996). This is where goods are a factor of production. For example, final goods may be produced from a variety of manufactured intermediate goods (Yamamoto 2003), manufactured goods may be produced using a variety of manufactured goods (which have not been consumed), and/or the innovation sector could use manufactured goods as a factor of production (Martin and Ottaviano 1999).

If the vertical linkage is in the innovation sector, this generates a feedback between growth and agglomeration with a similar result to localized knowledge spillovers. Martin and Ottaviano (1999) do not use the localized knowledge spillover mechanism demonstrated here, and instead, their innovation sector uses manufacturing goods as an innovation input such that the location of manufacturing affects the

cost of innovation through trade costs. Similarly, Yamamoto (2003) describes a model where final goods and innovation are produced using manufactured intermediate goods. This creates circular causation in growth and agglomeration because of the vertical linkages between intermediates and innovation.

12.5.3 Other Characteristics

There are many different factors which affect firm location decisions and subsequently space, innovation, and economic growth. Above, we have explored how these are dealt with in NEGG models by combining endogenous growth with the new economic geography and recognizing localized knowledge spillovers. But there are many more modeling choices for spatial factors which influence growth. For example, studying heterogeneous firms (Baldwin and Forslid 2010) helps describe the characteristics of which firms choose to locate in core or lagging regions. Other models include land requirements and continuous space (Desmet and Rossi-Hansberg 2009), whereby every firm is in a different location but willing to pay higher land rents to access more valuable locations. All of these have some influence on location choices for firms but ultimately demonstrate the same role of space in growth – that space is a barrier to knowledge transfer and technology diffusion which are inputs to innovation – and that policies or decisions by firms that reduce these spatial costs are growth enhancing.

12.6 Conclusions

We have described how NEGG models incorporate space into the Grossman and Helpman (1991a) product variety model of endogenous growth. Incorporating space into endogenous growth increases the complexity of these theoretical models. In all of these models with full local knowledge spillovers and partial global knowledge spillovers, space affects growth and growth affects location. The circular linked causality reinforces the core-periphery outcome of the NEG models. We show that integration between regions is more complex than is described by international trade models. In particular, we find that the cost of transferring knowledge between locations is important for firm location, stability, innovation, and growth.

From our discussion of the effect of space on growth through freeness of trade, agglomeration, and knowledge spillovers, there are a number of implications for economic policy in different locations. Agglomeration, freeness of trade, and knowledge spillovers are generally growth enhancing. The natural conclusion is that closer integration of economies will lead to increased growth rates. However, in these spatial models of growth, integration has two dimensions: trade costs and knowledge spillovers.

While traditional conceptions of integration refer to lowering of the cost of trading goods, Baldwin and Forslid (2000) show that combining theories of growth and space produces a more subtle view of integration where we can also view

integration as lowering the cost of trading information. Integration policies which focus solely on free trade may be destabilizing and result in a deindustrialization of the periphery region. Alternatively, integration policies that also focus on knowledge spillovers (or entirely on knowledge spillovers) will be growth enhancing for both regions. The model here shows how this form of integration is stabilizing, while pure trade cost integration can be destabilizing.

While lowering trade costs induces uneven development, it also results in higher rates of economic growth. Alternatively, policies that improve knowledge spillovers improve stability of the location of economic activity. Growth policies should consider the effect of trade, knowledge spillovers, labor, and capital market integration.

Acknowledgments We would like to thank Jacques Poot for excellent comments on earlier drafts of this chapter. Steven Bond-Smith would also like to thank the Royal Society of New Zealand Marsden Fund and the University of Waikato for financial support.

References

- Acs ZJ, Varga A (2002) Geography, endogenous growth, and innovation. *Int Reg Sci Rev* 25(1):132–148
- Aghion P, Howitt P (1992) A model of growth through creative destruction. *Econometrica* 60(2):132–148
- Audretsch DB, Feldman MP (1996) R&D spillovers and the geography of innovation and production. *Am Econ Rev* 86(3):630–640
- Baldwin RE, Forslid R (2000) The core-periphery model and endogenous growth: stabilizing and destabilizing integration. *Economica* 67(267):307–324
- Baldwin RE, Forslid R (2010) Trade liberalisation with heterogeneous firms. *Rev Dev Econ* 14(2):161–176
- Baldwin R, Martin P (2004) Agglomeration and regional growth. In: Henderson JV, Thisse J-F (eds) *Handbook of regional and urban economics*, vol 4, Cities and Geography. Elsevier, Amsterdam, pp 2671–2711
- Baldwin RE, Martin P, Ottaviano GI (2001) Global income divergence, trade and industrialisation: the geography of growth take-offs. *J Econ Growth* 6(1):5–37
- Baldwin R, Forslid R, Martin P, Ottaviano G, Robert-Nicoud R (2003) Economic geography and public policy. Princeton University Press, Princeton
- Desmet K, Rossi-Hansberg E (2009) Spatial growth and industry age. *J Econ Theory* 144(6):2477–2502
- Dixit AK, Stiglitz J (1977) Monopolistic competition and optimum product diversity. *Am Econ Rev* 67(3):297–308
- Eaton J, Kortum S (1999) International technology diffusion: theory and measurement. *Int Econ Rev* 40(3):537–570
- Faggian A, McCann P (2009) Human capital, graduate migration and innovation in British Regions. *Camb J Econ* 33(2):317–333
- Fujita M, Thisse J-F (2003) Does geographical agglomeration foster economic growth? And who gains and loses from it? *Jpn Econ Rev* 54(2):121–145
- Fujita M, Krugman P, Venables A (1999) *The spatial economy; cities, regions and international trade*. MIT Press, Cambridge, MA
- Grossman G, Helpman E (1991a) Innovation and growth in the global economy. MIT Press, Cambridge, MA

- Grossman G, Helpman E (1991b) Trade, knowledge spillovers, and growth. *Eur Econ Rev* 35(2–3):517–526
- Grossman G, Helpman E (1995) Technology and trade. In: Grossman G, Rogoff K (eds) *The handbook of international economics*, vol 3. Elsevier, North Holland, pp 1279–1337
- Kaldor N (1970) The case for regional policies. *Scott J Polit Econ* 17(3):337–348
- Krugman P (1979) Increasing returns, monopolistic competition and international trade. *J Int Econ* 9(4):469–479
- Krugman P (1991) Increasing returns and economic geography. *J Polit Econ* 99(3):483–499
- Krugman P, Venables A (1995) Globalisation and the inequality of nations. *Q J Econ* 110(4):857–880
- Lucas R (1988) On the mechanics of economic development. *J Monet Econ* 22(1):3–42
- Martin P, Ottaviano G (1999) Growing locations: industry location in a model of endogenous growth. *Eur Econ Rev* 43(2):281–302
- McCann P (2007) Sketching out a model of innovation, face-to-face interaction and economic geography. *Spat Econ Anal* 2(2):117–134
- McCann P (2009) Economic geography, globalization and New Zealand's productivity paradox. *N Z Econ Pap* 43(3):279–314
- Nelson RR (ed) (1993) National innovation systems: a comparative analysis. Oxford University Press, New York
- Romer P (1990) Endogenous technical change. *J Polit Econ* 98(5):S71–S102
- Venables A (1996) Equilibrium locations of vertically linked industries. *Int Econ Rev* 37(2):341–359
- Yamamoto K (2003) Agglomeration and growth with innovation in the intermediate goods sector. *Reg Sci Urban Econ* 33(3):335–360

Computable Models of Static and Dynamic Spatial Oligopoly

13

Amir H. Meimand and Terry L. Friesz

Contents

13.1	Introduction	238
13.2	The Notion of a Nash Equilibrium	239
13.3	Aspatial Oligopoly	239
13.3.1	Spatial Oligopoly	241
13.3.2	Variational Inequality (VI) Formulations of Spatial Oligopolist Competition	242
13.3.3	Diagonalization Algorithm for Variational Inequality	243
13.4	Static Network Oligopoly	244
13.5	Dynamic Network Oligopoly	246
13.5.1	Notation	246
13.5.2	The Firm's Objective Functional, Dynamics, and Constraints	246
13.5.3	The DVI Formulation of Dynamic Network Oligopoly	249
13.5.4	Discrete-Time Approximation	251
13.5.5	A Comment About Path Variables	251
13.5.6	Numerical Example	251
13.5.7	Interpretation of Numerical Results	254
13.6	Conclusions	257
	References	257

Abstract

Oligopolies are a fundamental economic market structure in which the number of competing firms is sufficiently small so that the profit of each firm is dependent upon the interaction of the strategies of all firms. There are alternative behavioral assumptions one may employ in forming a model of spatial oligopoly. In this chapter, we study the classical oligopoly problem based on Cournot's theory.

A.H. Meimand • T.L. Friesz (✉)

Department of Industrial and Manufacturing Engineering, Pennsylvania State University,

University Park, PA, USA

e-mail: amir@psu.edu; tlf13@psu.edu

The Cournot-Nash solution of oligopoly models assumes that firms choose their strategy simultaneously and each firm maximizes their utility function while assuming their competitor's strategy is fixed. We begin this chapter with the basic definition of Nash equilibrium and the formulation of static spatial and network oligopoly models as variational inequality (VI) which can be solved by several numerical methods that exist in the literature. We then move on to dynamic oligopoly network models and show that the differential Nash game describing dynamic oligopolistic network competition may be articulated as a differential variational inequality (DVI) involving both control and state variables. Finite-dimensional time discretization is employed to approximate the model as a mathematical program which may be solved by the multi-start global optimization scheme found in the off-the-shelf software package GAMS when used in conjunction with the commercial solver MINOS. We also present a small-scale numerical example for a dynamic oligopolistic network.

13.1 Introduction

The theory of oligopolistic markets and competition is introduced in Greenhut et al. (1987) and Greenhut and Lane (1989). Some basic models of spatial oligopoly are presented by Raa (1984), Dafermos and Nagurney (1987), Henderson and Quandt (1980), Novshek (1980), and Matsushima and Matsumura (2003). Moreover, Harker (1984), Dafermos and Nagurney (1987), and Nagurney (1999) studied the variational inequality (VI) approach to determine market equilibrium for a general static oligopoly model.

The motivation for this chapter is to construct computable general equilibrium models for static and dynamic spatial oligopoly with emphasis on computation rather than theory. We begin this chapter with the definition of Nash equilibrium and formulation of static spatial and network oligopoly as a variational inequality. In the network case, a few firms compete as Nash agents in the output market for a single homogeneous commodity. The firms are located at nodes of a transportation network in which common freight tariffs expressed as a fee per unit of flow on each arc are known and faced by each oligopolist. We then move on to computable modeling of dynamic network oligopoly described by differential Nash equilibria formulated as a differential variational inequality (DVI).

Models like those presented in this chapter arise when constructing spatial computable general equilibrium models, examples of which are Tobin and Friesz (1983), Friesz and Harker (1984), Dafermos and Nagurney (1987), Beckmann and Puu (1990), and Friesz (1993), as well as partial equilibrium models, when detailed freight flows are needed for a specific application. Throughout, our emphasis is on computation rather than theory, and our style is that of a simple primer in order to make the material accessible by the widest possible audience. Although some of the network models considered herein are notationally complex, our presentation relies only on very basic notions from microeconomic theory and elementary optimization.

13.2 The Notion of a Nash Equilibrium

Nash (1950, 1951) generalized the concept of equilibrium for the behavioral model consisting of N players who cannot improve their own self-interest by deviating from their equilibrium strategy, given that the other players use their equilibrium strategies (Friedman 1979). Suppose there are N players in a game and each player $i \in [1, N] \subseteq I_{++}$ (I_{++} being the set of positive integers) chooses a feasible strategy tuple x^i from the strategy set Ω_i to maximize the utility function $\Theta_i : \Omega_i \rightarrow \mathbb{R}^1$ (where \mathbb{R}^n refers to the real n -dimensional space), which will generally depend on other players' strategies. Every player $i \in [1, N]$ is trying to solve their best response problem:

$$\begin{aligned} \max & \quad \Theta(x^i; x^{-i}) \\ \text{subject to} & \quad x^i \in \Omega_i \end{aligned} \tag{13.1}$$

Note that, in Eq. (13.1), we use the notation

$$x^{-i} = (x^j, j \neq i) \quad i \in [1, N]$$

to refer to the tuple of strategies of players other than i . It is assumed that the non-own tuple x^{-i} is known to player i . A Nash equilibrium is the vector of strategies

$$x = (x^i : i \in [1, N])$$

formed from player-specific tuples such that each x^i solves the mathematical program (13.1). We denote a Nash equilibrium by $NE(\Theta, \Omega)$ (Friesz 2010).

13.3 Aspatial Oligopoly

Market structures are characterized by the existence of either a single or multiple firms. The former case, where the firm has no competitor, is referred to as a monopoly. Thus, he/she need not be concerned about the influence his/her strategy may have on competitors. The second case is known as an oligopoly, where the number of competing firms is sufficiently small so that the actions of any individual have a perceptible influence on the other firms. This interdependence of the firms' actions is the essential feature of oligopoly markets (Henderson and Quandt 1980).

Suppose there are N firms $i = 1, 2, \dots, N$ who supply a homogeneous commodity in a noncooperative fashion. Let $\pi(Q) : \mathbb{R}_+^1 \rightarrow \mathbb{R}_+^1$ (where \mathbb{R}_+^n is the set of

n tuples whose elements are real and nonnegative) denote the inverse demand function, where $Q \geq 0$ is the total supply in the market and $q^i \geq 0$ is the supply for firm i :

$$Q = \sum_{i=1}^N q^i \quad (13.2)$$

Further, let $V^i(q^i) : \mathbb{R}_+^1 \rightarrow \mathbb{R}_+^1$ denote the total production cost of firm i to supply q^i units. Given the other firms' strategy, q^{-i} , firm i tries to maximize its total profit $\Theta_i(q^i; q^{-i}) : \mathbb{R}_+^N \rightarrow \mathbb{R}^1$ by choosing the production quantity, q^i . To do so, firm i must solve the following mathematical program known as firm i 's best response problem:

$$\max_{q_i} \Theta_i(q^i; q^{-i}) = q^i \pi(Q) - V^i(q^i) \quad (13.3)$$

subject to

$$\begin{aligned} Q &= \sum_{i=1}^N q^i \\ q^i &\geq 0 \end{aligned}$$

There is a Nash game between the players (firms) choosing a feasible strategy $q^i \geq 0$ where every player's objective function depends on all other players' strategies. The strategy $q^* = (q^{*1}, q^{*2}, \dots, q^{*N}) \in \mathbb{R}_+^N$ is the solution of this game, and, in general, we refer to the solution of a Nash game as a Nash equilibrium, $NE(\Theta, \Omega)$ (Harker 1984).

If $V^i(\cdot)$ is convex and continuously differentiable for $i = 1, 2, \dots, N$, the inverse demand function $\pi(\cdot)$ is strictly decreasing and continuously differentiable, and the industry revenue function $Q^T \pi(Q)$ is concave, then $q^* = (q^{*1}, q^{*2}, \dots, q^{*N})$ is a Nash equilibrium solution if and only if

- (i) $[\pi(Q^*) + q^{i*} \nabla_{q^i} \pi(Q^*) - \nabla_{q^i} V^i(q^{i*})] q^{i*} = 0 \quad \forall i = 1, 2, \dots, N$
- (ii) $\pi(Q^*) + q^{i*} \nabla_{q^i} \pi(Q^*) - \nabla_{q^i} V^i(q^{i*}) \leq 0 \quad \forall i = 1, 2, \dots, N$

where

$$\begin{aligned} Q^* &= \sum_{i=1}^N q^{i*} \\ q^{i*} &\geq 0 \end{aligned}$$

Conditions (i) and (ii) are the first-order necessary (and sufficient) conditions for the set of problems defined by Eq. (13.3) for each $i = 1, 2, \dots, N$ (Murphy et al. 1982).

13.3.1 Spatial Oligopoly

Spatial models try to determine the factors governing the distribution and/or location of economic activity over space. In spatial models, space is often represented as a set of discrete locations at a certain level of resolution.

We may now generalize the classical model of oligopoly to the spatial case in which a few firms are competing in spatially separated markets. Assume there are N firms and M demand markets that are generally spatially separated. Further, assume a homogenous commodity is produced by N firms and consumed in M markets. Let q^i denote the output of firm $i \in [1, N] \subseteq I_{++}$ and c^j denote the demand for the commodity at market $j \in [1, M] \subseteq I_{++}$. Let s_j^i denote the commodity shipment from firm i to market j . The following conservation of flow equations for the total supply and demand must hold

$$q^i - \sum_{j=1}^M s_j^i = 0 \quad \forall i = 1, 2, \dots, N \quad (13.4)$$

$$c^j - \sum_{i=1}^N s_j^i = 0 \quad \forall j = 1, 2, \dots, M \quad (13.5)$$

where

$$0 \leq c \leq c^{\max}, 0 \leq q \leq q^{\max}, 0 \leq s \leq s^{\max} \quad (13.6)$$

Note that in Eq. (13.6), these column vectors are employed:

$$\begin{aligned} q &\in \Re_{++}^N \\ c &\in \Re_{++}^M \\ s &\in \Re_{++}^{N \times M} \end{aligned}$$

denote the production outputs, demands, and the commodity shipments. As previously, $V^i(q^i)$ is the total production cost for firm i to supply q^i units, r_j^i is the freight rate (tariff) charged per unit of flow s_j^i , and $\pi_j(c^j)$ is the inverse demand function of market j . To consider the general situation, we assume the production cost may depend on the entire production pattern $V^i = V^i(q) : \Re_{++}^N \rightarrow \Re_{++}^1$, the inverse demand

function at market j may depend on the entire consumption pattern, $\pi_j = \pi_j(c) : \Re_{++}^M \rightarrow \Re_{++}^1$, and the freight rate (tariff) may depend on the entire shipment pattern, $r_j^i = r_j^i(s) : \Re_{++}^{N \times M} \rightarrow \Re_{++}^1$ (where \Re_{++}^n is the set of n tuples whose elements are real and positive). The aforementioned dependency means the transportation cost functions allow for interactions among the firms' shipments (as would be the case when the transportation system has limited capacity and the firms all use the same system). In this model, each of the firms wants to maximize their profit. The profit function of each firm i may be expressed as follows (Dafermos and Nagurney 1987):

$$\Theta_i(c, q, s) = \sum_{j=0}^M \pi_j(c) s_j^i - V^i(q) - \sum_{j=0}^M r_j^i(s) s_j^i \quad (13.7)$$

In light of Eqs. (13.4) and (13.5), one may write

$$\Theta = \Theta(s)$$

Now we consider the usual market mechanism associated with an oligopoly, for which the N firms behave as noncooperative agents and supply the commodity while maximizing their own profit. We will seek a commodity shipment pattern s for which the N firms will be in a state of equilibrium per the following definition (Nagurney 1999):

Definition i

(spatial Cournot-Nash equilibrium) A commodity shipment pattern s^* is said to constitute a Cournot-Nash equilibrium if for each $i = 1, 2, \dots, N$,

$$\Theta_i(s_j^{i*}, s_j^{-i*}) \geq \Theta_i(s_j^i, s_j^{-i*}) \quad \forall s_j^i \in \Omega$$

where

$$\Omega = \{s_j^i : (4), (5), (6)\}$$

So, firm i can maximize its profit, Θ_i , by choosing s_j^{i*} among all other possible shipment patterns when the shipment patterns of the other competitors, s_j^{-i*} , are known and fixed.

13.3.2 Variational Inequality (VI) Formulations of Spatial Oligopolist Competition

Variational inequalities have been used to study equilibrium problems, but now are increasingly used to formulate a much wider array of mathematical problems.

In this section, we present the variational inequality equivalent to a Cournot-Nash equilibrium in the following theorem:

Theorem i

(*Nash equilibrium equivalent to a variational inequality*). *The general Nash equilibrium $\text{NE}(\Theta, \Omega)$ is equivalent to the following variational inequality $VI(\nabla\Theta, \Omega)$:*

$$\text{find } x^* \in \Omega$$

such that

$$[\nabla\Theta(x^*)]^T(x - x^*) \geq 0 \quad \forall x \in \Omega$$

the following regularity conditions hold: (i) each $\Theta^i(x) : \Omega_i \rightarrow \mathbb{R}^1$ is convex and continuously differentiable in x^i , and (ii) each Ω_i is a closed convex subset of \mathbb{R}^n . See Friesz (2010) for a proof of the above result.

In geometric terms, the variational inequality states that $[\nabla\Theta(x^*)]^T$ is orthogonal to the feasible set Ω at the point x^* . Applying Theorem i, the oligopoly model can be formulated as the following VI (Nagurny 1999):

$$\text{find } q^* \geq 0 \text{ such that :}$$

$$\sum_{i=1}^N [\nabla_{q^i} V^i(q^{i*}) - \pi(Q^*) - \nabla_{q^i} \pi(Q^*) q^{i*}] (q^i - q^{i*}) \geq 0 \quad \forall q \geq 0 \quad (13.8)$$

where

$$Q^* = \sum_{i=1}^N q^{i*}$$

The VI formulation (13.8) states that if each firm i chooses strategy q^{i*} , no firm can increase their profit by changing their strategy, while the other players' strategies remain unchanged.

13.3.3 Diagonalization Algorithm for Variational Inequality

In this section, the diagonalization algorithm (or diagonalization for short) discussed is one of several algorithms that can be used to solve a variational inequality. Its algorithmic philosophy is very similar to the Gauss-Seidel method from the numerical analysis literature (Friesz 2010). Diagonalization is suitable for solving finite-dimensional variational inequalities because the resulting subproblems are nonlinear

programs that can be efficiently solved with well-understood nonlinear programming algorithms, which are often available in the form of commercial software. This fact notwithstanding, diagonalization may fail to converge, and its use on large-scale problems can be frustrating. We are now ready to state the so-called *diagonalization algorithm* to solve $VI(F, \Omega)$. The algorithm is composed of three main steps:

Step 0. Initialization. Determine a feasible solution $x^0 \in \Omega$ and set $k = 0$.

Step 1. Solve diagonalized variational inequality. Form the separable functions

$F_i^k(x_i)$ for all $i \in [1, N] \subseteq I_{++}$ and solve the associated diagonalized variational inequality problem. That is, find $x^{k+1} \in \Omega$ such that

$$\sum_{i=1}^N F_i(x_i^{k+1})(x_i - x_i^{k+1}) \leq 0 \quad \forall x \in \Omega$$

which may be solved using the following nonlinear program:

$$\max J(x) = \sum_{i=1}^N \int_0^{x_i} F_i^k(z_i) dz_i \text{ s.t. } \forall x \in \Omega$$

where z_i are dummy variables of integration.

Step 2. Stopping criteria. For a small positive number $\xi \in \mathbb{R}_{++}^1$, a preset tolerance, if

$$\max_{i \in [1, N]} |x_i^k - x_i^{k+1}| < \xi$$

stop; otherwise, set $k = k + 1$ and go to Step 1.

Other numerical methods for VI such as gap function method, fixed-point method, and successive linearization and Lemke's algorithm are presented in Friesz (2010) with some numerical examples.

13.4 Static Network Oligopoly

A transportation network allows firms to ship their commodity to multiple markets through multiple paths. In this section, we consider the impact of spatial transportation networks on the firms' economic decisions. For the time being, we restrict our consideration to the static case, while in a later section, we will extend this model to the dynamic case. Several papers exist in the literature about modeling spatial network oligopoly. For example, Hakimi (1983) chooses locations for new facilities in an oligopolistic environment. Tobin and Friesz (1983) show that an equivalent optimization problem for spatial price network equilibrium may be formulated without path variables. Moreover, Rovinskey et al. (1980) developed a model for Cournot-Nash equilibrium of firms who are spatially separated but supply a commodity to a single market. Hashimoto (1985) and Harker (1986) formulated more general Cournot-Nash models in which consumers are also spatially dispersed.

In this section, we are going to present the static network oligopoly model. Suppose we have a transportation network with several spatially separated markets for a particular commodity located at nodes within a network. Each market may be supplied by several firms with production facilities located at nodes within the network (these nodes may or may not be the same nodes as a market). Further, a firm may have production facilities at more than one node. Each firm is seeking to maximize its total profit by making decisions about their output rate, allocation of this output to each market, and shipment pattern between OD (origin–destination) pairs. So, there is a Nash game among firms whose facilities and final demand markets are fixed at distinct nodes of a distribution network and connected by paths involving chains of arcs of that network. The network is represented by $G(N, W)$, in which N is a set of nodes and W is a set of OD pairs (i, j) for which there exists at least one path from i to j . Further, F is a set of firms that compete in the network. N_f is the set of nodes at which firm f has economic presence, and W_f is the set of OD pairs used by firm f to transport the commodity. Moreover, s_{ij}^f is the shipment of firm f for OD pairs (i, j) , r_{ij} is the freight rate (tariff) charged per unit of flow for OD pair (i, j) , c_i^f is the allocation of output of firm $f \in F$ at node $i \in N_f$ to consumers, q_i^f is the output of firm $f \in F$ at node $i \in N_f$, and $V_i^f(q_i^f)$ is the variable cost of production for q_i^f units. The best response problem for firm f is

$$\max \Theta_f(c^f, q^f, h^f; c^{-f}, q^{-f}, h^{-f}) = \sum_{i \in N_f} \pi_i \left(\sum_{g \in F} c_i^g \right) c_i^f - \sum_{i \in N_f} V_i^f(q_i^f) - \sum_{(i,j) \in W_f} r_{ij} s_{ij}^f$$

subject to

$$c_i^f = q_i^f + \sum_{(i,j) \in W} s_{ji}^f - \sum_{(j,i) \in W} s_{ij}^f \quad \forall i \in N_f \quad (13.9)$$

$$0 \leq c_i^f \leq c_i^{\max}, 0 \leq q_i^f \leq q_i^{\max}, 0 \leq s_{ij}^f \leq s_{ij}^{\max} \quad (13.10)$$

In keeping with the above, each firm maximizes its total profit $\Theta_f(c^f, q^f, h^f; c^{-f}, q^{-f}, h^{-f})$. Constraint (13.9) guarantees that flow is conserved at each node. Moreover, by Eq. (13.10), all consumption, production, and shipping variables are nonnegative and bounded from above. Applying Theorem i, the static network oligopoly can be formulated as the following VI:

$$\begin{aligned} & \text{find } (c_i^{f*}, q_i^{f*}, s_{ij}^{f*}) \in \Omega \text{ such that} \\ & \sum_{f \in F} \left[\sum_{i \in N_f} \nabla_{q_i^f} \Theta_f(q_i^f - q_i^{f*}) + \sum_{i \in N_f} \nabla_{c_i^f} \Theta_f(c_i^f - c_i^{f*}) + \sum_{(i,j) \in W_f} \nabla_{s_{ij}^f} \Theta_f(s_{ij}^f - s_{ij}^{f*}) \right] \geq 0, \forall (c, q, s) \in \Omega \end{aligned} \quad (13.11)$$

where

$$\Omega = \{(c_i^f, q_i^f, s_i^f) : \text{Eqs. (13.9), (13.10)}\}$$

13.5 Dynamic Network Oligopoly

We now turn to the problem of modeling and computing differential Nash equilibria among the oligopolistic firms. The oligopolistic firms of interest, embedded in a network economy, are in oligopolistic competition according to dynamics that describe the trajectories of inventories/backorders and correspond to flow conservation for each firm at each node of the network. The oligopolistic firms, acting as shippers, perfectly compete as price takers in the market for physical distribution services. Perfect competition in shipping arises because numerous shipping companies serve numerous customers due to the involvement of shippers in the numerous output markets of the network economy. The time scale we consider is neither short nor long, but rather of sufficient length to allow output and shipping pattern adjustments, while not long enough for firms to relocate, enter, or leave the network economy. This model, which takes the form of a differential variational inequality, is presented in this section. Dynamic network oligopoly models were studied by Brander and Zhang (1993), Nagurney et al. (1994), Wie and Tobin (1997), Nagurney et al. (2002), Friesz et al. (2006), and Markovich (2008). To develop the mathematical formulation of network oligopoly, we follow the exposition in Friesz (2010).

13.5.1 Notation

Fortunately, much of the notation introduced in previous sections of this chapter is still relevant. Yet, because there are some subtle differences between the dynamic oligopoly model that we now study and problems explored previously in this chapter, we choose to give an exhaustive list of the notation to be employed below, even though that will involve some duplication. We let continuous time be denoted by the scalar $t \in \mathbb{R}_+^1$, initial time by $t_0 \in \mathbb{R}_+^1$, and final time by $t_f \in \mathbb{R}_{++}^1$, with $t_0 < t_f$ so that $t \in [t_0, t_f] \subset \mathbb{R}_+^1$. There are several sets important to articulating a network oligopoly model; these are as follows: \mathcal{F} for firms, \mathcal{A} for directed arcs, \mathcal{N} for nodes, and \mathcal{W} for origin–destination (OD) pairs. Subsets of these sets are formed as is meaningful by using the subscripts f for a specific firm, i for a specific node, and ij for a specific OD pair (i, j) .

Each firm $f \in \mathcal{F}$ controls production (output) rates q^f , allocation of output to meet demand c^f , and shipping pattern s^f . Inventories I^f are state variables determined by the controls. In particular, concatenations of the firm-specific vectors c^f , q^f , and s^f for the vectors c , q , and s , respectively.

13.5.2 The Firm's Objective Functional, Dynamics, and Constraints

Each firm has the objective of maximizing net profit expressed as revenue less cost and takes the form of an operator acting on allocations of output to meet demands, production rates, and shipment patterns. For each firm $f \in \mathcal{F}$, net profit is given by the following functional:

$$\begin{aligned}
\Phi_f(c^f, q^f, s^f; c^{-f}, q^{-f}) = & e^{-\rho t_f} Z_f[I(t_f), t_f] \\
& + \int_{t_0}^{t_f} e^{-\rho t} \left\{ \sum_{i \in \mathcal{N}} \pi_i \left(\sum_{g \in \mathcal{F}} c_i^g, t \right) c_i^f - \sum_{i \in \mathcal{N}_f} V_i^f(q^f, t) \right. \\
& - \sum_{(i,j) \in \mathcal{W}_f} r_{ij}(t) s_{ij}^f - \sum_{i \in \mathcal{N}_f} V_i^f(q^f, t) - \sum_{(i,j) \in \mathcal{W}_f} r_{ij}(t) s_{ij}^f \\
& \left. - \sum_{i \in \mathcal{N}} \psi_i^f(I_i^f, t) \right\} dt
\end{aligned} \tag{13.12}$$

where $\rho \in \Re_{++}^1$ is a constant nominal rate of discount, $r_{ij} \in \Re_{++}^1$ is the freight rate (tariff) charged per unit of flow s_{ij} for OD pair $(i, j) \in \mathcal{W}_f$, ψ_i^f is firm f 's inventory cost at node i , and I_i^f is the inventory/backorder of firm f at node i . In Eq. (13.12), c_i^f is the allocation of the output of firm $f \in \mathcal{F}$ at node $i \in \mathcal{N}$ to consumption at that node. Also $Z_f[I^f(t_f), t_f]$ is the liquidation value of inventory remaining at the terminal time, where $I^f = (I_i^f : i \in \mathcal{N}_f)$. Because our formulation is in terms of flows, it is convenient to employ the inverse demand functions $\pi_i(c_i, t)$ where $c_i = \sum_{g \in \mathcal{F}} c_i^g$ is the total allocation of output to consumption for node i . Furthermore, q_i^f is the output of firm $f \in \mathcal{F}$ at node $i \in \mathcal{N}$. Again $V_i^f(q, t)$ is the variable production cost for firm $f \in \mathcal{F}$ at node $i \in \mathcal{N}$. The reader should note that $\Phi_f(c^f, q^f, s^f; c^{-f}, q^{-f})$ is a functional that is determined by the controls c^f , q^f , and s^f when non-own allocations to consumption and non-own production rates

$$c^{-f} \equiv (c^f : f' \neq f), q^{-f} \equiv (q^f : f' \neq f)$$

are taken to be exogenous data by firm f . The first term of the objective functional $\Phi_f(c^f, q^f, s^f; c^{-f}, q^{-f})$ in expression (13.12) is the firm's revenue, the second term is the firm's cost of production, the third term is the firm's shipping costs, and the last term is the firm's inventory or holding cost.

We next impose the terminal time inventory constraints

$$I_i^f(t_f) \geq \tilde{K}_i^f \quad \forall f \in \mathcal{F}, i \in \mathcal{N}_f \tag{13.13}$$

where $\tilde{K}_i^f \in \Re_{++}^1$ are exogenous. Again all consumption, production, and shipping variables are nonnegative and bounded from above; that is,

$$C^f \geq c^f \geq 0 \tag{13.14}$$

$$Q^f \geq q^f \geq 0 \tag{13.15}$$

$$S^f \geq s^f \geq 0 \tag{13.16}$$

where

$$C^f \in \Re_{++}^{|\mathcal{F}|}, Q^f \in \Re_{++}^{|\mathcal{F}|}, S^f \in \Re_{++}^{|\mathcal{W}_f|}$$

As for the monopoly, constraints (13.14), (13.15), and (13.16) are recognized as pure control constraints, while (13.1) are terminal conditions for the state space variables. Naturally

$$\Omega_f = \{(c^f, q^f, s^f) : \text{Eqs. (13.14), (13.15), (13.16)}\}$$

is the set of feasible controls.

Firm f solves an optimal control problem to determine its production q^f , allocation of production to meet demand c^f , and shipping pattern s^f by maximizing its profit functional $\Phi_f(c^f, q^f, s^f; c^{-f}, q^{-f})$ subject to inventory dynamics expressed as flow balance equations and pertinent production and inventory constraints. The inventory dynamics for firm $f \in \mathcal{F}$, expressing simple flow conservation, obey

$$\frac{dI_i^f}{dt} = q_i^f + \sum_{(j,i) \in \mathcal{W}} s_{ji}^f - \sum_{(i,j) \in \mathcal{W}} s_{ij}^f - c_i^f \quad \forall i \in \mathcal{N}_f \quad (13.17)$$

$$I_i^f(t_0) = K_i^f \quad \forall i \in \mathcal{N}_f \quad (13.18)$$

$$I_i^f(t_f) = \tilde{K}_i^f \quad \forall i \in \mathcal{N}_f \quad (13.19)$$

where $K_i^f \in \Re_{++}^1$ and $\tilde{K}_i^f \in \Re_+^1$ are exogenous. In addition to the terminal time inventory (state) constraints (13.19), the model is general enough to handle inventory constraints over the entire planning horizon $[t_0, t_f]$. For instance, nonnegativity of the inventory (state) variables could be imposed to restrict firms from taking backorders.

We next note that firm f 's problem is with the c^{-f} and q^{-f} as exogenous inputs; compute c^f , q^f , and s^f (thereby finding I^f) in order to solve the following extremal problem:

$$\max_{\substack{\text{s.t.} \\ (c^f, q^f, s^f) \in \Omega_f}} \left. \Phi_f(c^f, q^f, s^f; c^{-f}, q^{-f}) \right\} \forall f \in \mathcal{F} \quad (13.20)$$

where

$$\Omega_f = \{(c^f, q^f, s^f) : \text{Eqs. (13.13), (13.14), (13.15), (13.16), (13.17), (13.18), (13.19) hold}\}$$

also for all $f \in \mathcal{F}$. That is, each firm is a Nash agent that knows and employs the current instantaneous values of the decision variables of other firms to make its own noncooperative decisions. As such, Eq. (13.20) is a differential Nash game. Moreover, each firm's best response problem (13.20) is a continuous time optimal control problem.

13.5.3 The DVI Formulation of Dynamic Network Oligopoly

To continue our discussion of oligopoly, we assume the Nash game expressed above is regular in the sense of the following definition:

Definition ii

The dynamic oligopolistic network competition problem introduced above will be considered regular if (i) the state operator $I(c, q, s)$ exists and is unique, while each of its components is continuous and G differentiable; (ii) the inverse demand, production cost, and inventory cost functions are continuously differentiable with respect to controls and states; and (iii) for each $f \in \mathcal{F}$, the composite terminal cost function

$$Z_f[I^f(t_f), t_f] + \sum_{i \in \mathcal{N}_f} \gamma_i^f [\tilde{K}_i^f - I_i^f(t_f)]$$

is continuously differentiable with respect to $I_i^f(t_f)$ for all $i \in \mathcal{N}_f$.

In the above definition, each γ_i^f is a constant dual variable that prices out the terminal constraint on inventory. We also note that Eq. (13.20) is an optimal control problem with fixed terminal time. Also

$$\Psi_f(c^f, q^f, s^f, I^f, \alpha^f, \beta^f, \lambda^f) = \sum_{i \in \mathcal{N}_f} \lambda_i^f \left(q_i^f + \sum_{(j,i) \in \mathcal{W}} s_{ji}^f - \sum_{(i,j) \in \mathcal{W}} s_{ij}^f - c_i^f \right) \quad (13.21)$$

where $\alpha_i^f \in \Re_+^1$ and $\beta_i^f \in \Re_+^1$ are dual variables for the inventory-bounding constraints (13.13), while $\alpha^f \in \Re^{|\mathcal{N}_f|}$ and $\beta^f \in \Re^{|\mathcal{N}_f|}$; also $\lambda_i^f \in \Re_+^1$ is the adjoint variable for the dynamics of firm f at node i , while $\lambda^f \in (\mathcal{H}^1[t_0, t_f])^{|\mathcal{W}|}$. Clearly Φ_f is the instantaneous profit. To interpret Ψ_f , we need to understand the relevant dynamic shadow benefits and shadow costs of this model. To that end, recall that, along an optimal trajectory, the adjoint variables obey

$$\lambda_i^f = \frac{\partial J_f}{\partial I_i^f}$$

Consequently,

$$\Psi_f = \sum_{i \in \mathcal{N}_f} \frac{\partial J_f}{\partial I_i^f} \frac{dI_i^f}{dt}$$

which is recognized as the shadow value of dynamic benefits arising from current inventory held; it can be either a cost or a benefit, depending on its sign.

Familiarity with variational inequalities suggests that the following variational inequality has solutions that are differential Nash equilibria for a noncooperative game in which individual firms maximize net profits in light of current information about their competitors:

$$\begin{aligned} & \text{find } (c^{f*}, q^{f*}, s^{f*}) \in \Omega \text{ such that} \\ 0 \geq & \sum_{f \in \mathcal{F}} \int_{t_0}^{t_f} \left[\sum_{i \in \mathcal{N}_f} \frac{\partial \Phi_f^*}{\partial c_i^f} (c_i^f - c_i^{f*}) + \sum_{i \in \mathcal{N}_f} \frac{\partial \Phi_f^*}{\partial q_i^f} (q_i^f - q_i^{f*}) \right. \\ & \left. + \sum_{(i,j) \in \mathcal{W}_f} \frac{\partial \Phi_f^*}{\partial s_{ij}^f} (s_{ij}^f - s_{ij}^{f*}) \right] dt \quad \text{for all } (c, q, s) \in \Omega \end{aligned} \quad (13.22)$$

where

$$\Phi_f^* = \Phi_f(c^{f*}, q^{f*}, s^{f*}, I^{f*}; c^{-f}, q^{-f}; t) \quad (13.23)$$

$$\Omega = \prod_{f \in \mathcal{F}} \Omega_f \quad (13.24)$$

We note that Eq. (13.22) is a differential variational inequality expressing the differential Nash game that is our present interest. This formulation also provides guidance in devising a computational strategy, as we show in Sect. 13.5.4.

The issue of immediate concern is to formally demonstrate that solutions of Eq. (13.22) are differential Nash equilibria. In fact, we state and prove the following result:

Theorem ii

(*Differential variational inequality formulation of dynamic oligopolistic network competition.*) Any solution of Eq. (13.22) is a solution of the dynamic oligopolistic network competition problem when regularity in the sense of Definition i holds. See Friesz (2010) for a proof of the above result.

We next note that the following existence result holds:

Theorem iii

(*Existence of dynamic oligopolistic network equilibrium.*) When the variational inequality of Theorem ii is regular in the sense of Definition i, there exists a solution of the dynamic oligopolistic network competition problem. The preceding result is proven in Friesz (2010).

13.5.4 Discrete-Time Approximation

If we define the discrete instant of time $t_k = t_0 + k\Delta$, where Δ is the time step employed, while

$$N = \frac{t_f - t_0}{\Delta}$$

is the number of discretizations, and $t_N = t_f$. Then, a finite-dimensional mathematical program is created, which forms a computable approximate solution if its objective is convex.

13.5.5 A Comment About Path Variables

It should be noted that one may introduce path flows in the above formulation by reexpressing the state dynamics as

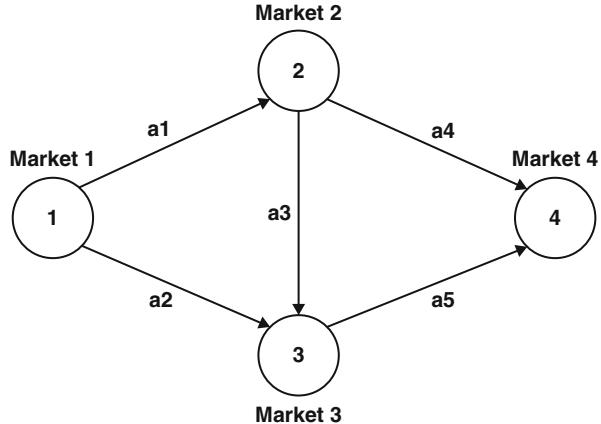
$$\frac{dI_i^f}{dt} = q_i^f + \sum_{j \in \mathcal{N}_f} \sum_{p \in P_{ji}} h_p^f - \sum_{j \in \mathcal{N}_f} \sum_{p \in P_{ij}} h_p^f - c_i^f$$

for every firm $f \in \mathcal{F}$ and node $i \in \mathcal{N}_f$, where P_{ji} is the set of paths from node $j \in \mathcal{N}_f$ to node $i \in \mathcal{N}_f$; furthermore, h_p is the flow on path $p \in P_{ji}$. There are corresponding, but quite obvious, changes in the firm's objective function and the upper and lower bound constraints on its controls. We omit a complete statement of such details for the sake of brevity.

13.5.6 Numerical Example

Let us consider a network of five arcs, four nodes, and four firms, where a single firm f is located at each node $i = 1, 2, 3, 4$. Consumption of each firm's output potentially occurs at every node; this consumption may be of local or of imported output as the network topology permits. Figure 13.1 illustrates the network.

Fig. 13.1 Network of five arcs, four nodes, and four firms



The time interval of interest is $[0, 20]$; that is, $t_0 = 0, t_f = 20$. In this example, firm 1 has an economic presence at all nodes; firm 2 at nodes 2, 3, and 4; firm 3 at nodes 3 and 4; and finally firm 4 at node 4 only. Therefore, $\mathcal{F} = \{1, 2, 3, 4\}$, $\mathcal{N}_1 = \{1, 2, 3, 4\}$, $\mathcal{N}_2 = \{2, 3, 4\}$, $\mathcal{N}_3 = \{3, 4\}$, and $\mathcal{N}_4 = \{4\}$. Before time discretization, there are 29 controls and ten state variables associated with this example; these are enumerated in Table 13.1:

At time $t_0 = 0$, every firm has an initial inventory of 100 units at their respective locations. That is, $I_i^f(0) = 100$ for $f \in \mathcal{F}$ and $i \in \mathcal{N}_f$. In addition, we impose the condition that no backordering is allowed by any firm at any node at the terminal time $t_f = 20$. That is,

$$I_i^f(20) \geq 0 \text{ for } f \in \mathcal{F} \text{ and } i \in \mathcal{N}_f \quad (13.25)$$

The inventory dynamics are the flow balance equations

$$\begin{aligned} \frac{dI_1^1}{dt} &= q_1^1 - h_1^1 - h_2^1 - h_3^1 - h_4^1 - h_5^1 - h_6^1 - c_1^1 \\ \frac{dI_2^1}{dt} &= h_1^1 - h_7^1 - h_8^1 - h_9^1 - c_2^1 \\ \frac{dI_3^1}{dt} &= h_2^1 + h_3^1 + h_7^1 - h_{10}^1 - c_3^1 \\ &\vdots \\ \frac{dI_4^4}{dt} &= q_4^4 - c_4^4 \end{aligned} \quad (13.26)$$

which we only partially enumerate in the interest of saving space. We assume the inverse demands at each node i take the following form:

$$\pi_i(c_i, t) = \alpha_i - \beta_i (c_i)^m \quad (13.27)$$

Table 13.1 State and control variables

Firm	Controls by node or path				States			
1	c_1^1	c_2^1	c_3^1	c_4^1		I_1^1	I_2^1	I_3^1
2		c_2^2	c_3^2	c_4^2		I_2^2	I_3^2	I_4^2
3			c_3^3	c_4^3			I_3^3	I_4^3
4				c_4^4				I_4^4
all	q_1^1	q_2^2	q_3^3	q_4^4				
1	h_1^1	h_2^1	h_3^1	h_4^1	h_5^1	h_6^1	h_7^1	h_8^1
2						h_7^2	h_8^2	h_9^2
3								h_{10}^3

Table 13.2 Path–arc

Path	Arc sequence
p_1	a_1
p_2	a_2
p_3	a_1, a_3
p_4	a_1, a_4
p_5	a_1, a_3, a_5
p_6	a_2, a_5
p_7	a_3
p_8	a_4
p_9	a_3, a_5
p_{10}	a_5

where $m \in \mathbb{R}_{++}^1$ is a constant. Also $\alpha_i \in \mathbb{R}_{++}^1$ and $\beta_i \in \mathbb{R}_{++}^1$ for all i are constants. The production cost functions of each firm f have the form

$$V_i^f = \frac{1}{2} \rho_i^f (q_i^f)^2 + \frac{1}{3} \sigma_i^f (q_i^f)^3 \text{ for all } i = 1, \dots, 4 \quad (13.28)$$

where ρ_i^f and $\sigma_i^f \in \mathbb{R}_{++}^1$ are also constants for all allowed i and f . In Eq. (13.28), we consider a nonconvex production cost functions in order to capture both increasing and decreasing economies of scale for different production rate regimes. We assume the holding costs are quadratic and of the form

$$\psi_i^f = \frac{1}{2} \eta_i^f (I_i^f)^2 \text{ for } f \in \mathcal{F} \text{ and } i \in \mathcal{N}_f \quad (13.29)$$

where $\eta_j^f \in \mathbb{R}_{++}^1$ are constants, again for all allowed i and f . The relationships between arc and path variables are summarized in Table 13.2.

Furthermore, the relevant arc-path incidence matrix is

$$\Delta_p = (\delta_{ap}) = \begin{bmatrix} 1 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 \end{bmatrix}$$

The associated path costs are

$$R = \Delta^T r \quad (13.30)$$

where

$$r = (r_{a_i} : i = 1, 2, 3, 4, 5)$$

and

$$r_{a_i} = A_i + B_{a_i}(f_i)^n \quad i = 1, 2, 3, 4, 5$$

are unit freight rates for individual arcs and the $A_i \in \Re_{++}^1$ and $B_i \in \Re_{++}^1$ are known constants. We impose the following vectors of bounds on control variables:

$$C^f = Q^f = H^f = 75$$

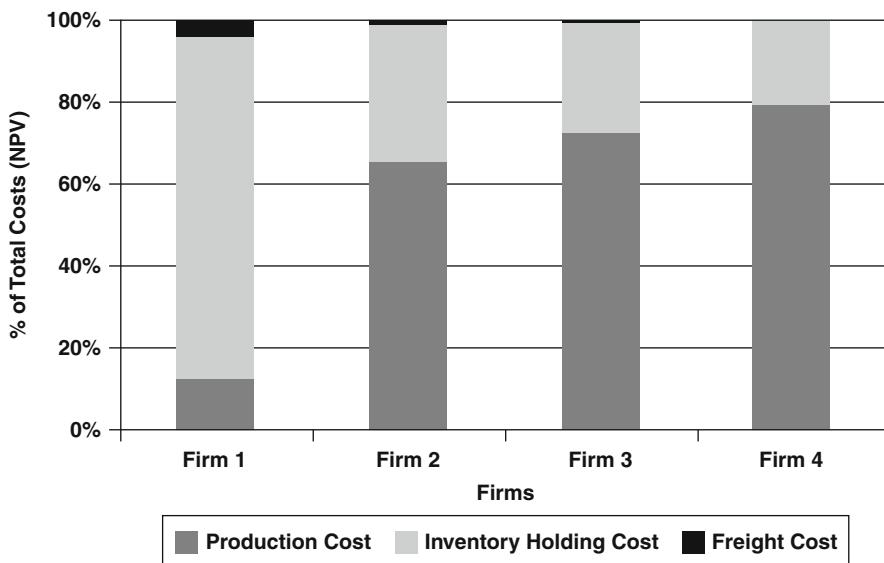
Each firm's instantaneous profit function is found by substituting Eqs. (13.27), (13.28), (13.29), and (13.30) into Eq. (13.12), where $\rho \in \Re_{++}^1$ is again the fixed nominal interest rate. A discrete-time approximation of the corresponding differential variational inequality is created using $N = 21$ equal time steps. The resulting finite-dimensional variational inequality is restated as a nonlinear complementarity problem and solved using GAMS with the PATH solver. The numerical values of the model's parameters are presented in the [Table 13.3](#):

13.5.7 Interpretation of Numerical Results

In [Fig. 13.2](#), we compare the net present of cumulative production, inventory holding, and transportation costs incurred by the four firms. The net present value of profit for firm 1 is $-\$185,592$, $-\$926,070$ for firm 2, $+\$248,179$ for firm 3, and $-\$314,978$ for firm 4. Thus, the only firm to realize a positive profit is firm 3, whereas all other firms experience losses. Further, [Fig. 13.3](#) shows that firm 1, which has an economic presence at all 4 nodes, has the highest transportation costs. By comparison, firms 2 and 3 have relatively small transportation costs, while firm 4 incurs no transportation cost. This is expected, since the economic presence of firms

Table 13.3 Model's parameters

Parameter	Value	Parameter	Value	Parameter	Value
ρ	0.05	A_1	2	A_2	2
A_3	2	A_4	2	A_5	2
B_1	0.9	B_2	0.9	B_3	0.9
B_4	0.9	B_5	0.9	α_1	2,000
β_1	12	α_2	2,200	β_2	16
α_3	2,400	β_3	14	α_4	2500
β_4	18	ρ_1^1	0.3	ρ_2^2, ρ_4^4	0.1
ρ_3^3	0.2	$\sigma_i^i, i = 1, \dots, 4$	1	$\eta_2^1, \eta_4^1, \eta_4^3, \eta_4^4$	1
η_1^1	4	$\eta_3^1, \eta_2^2, \eta_3^2$	2	η_4^2, η_3^3	3
t_0	0	t_f	20	N	20
Δ	1	n	1	m	1

**Fig. 13.2** Cost by firms

2 and 3 throughout the network is comparatively small, and firm 4 only has economic presence in its home market.

The production rates of the four firms and the prices of finished goods in the four markets are plotted in [Fig. 13.3](#). [Figure 13.3a](#) shows that in market 4, where all the firms can compete, the price increases significantly over time. In contrast, market 1, in which only firm 1 has economic presence, shows a relatively small change in price.

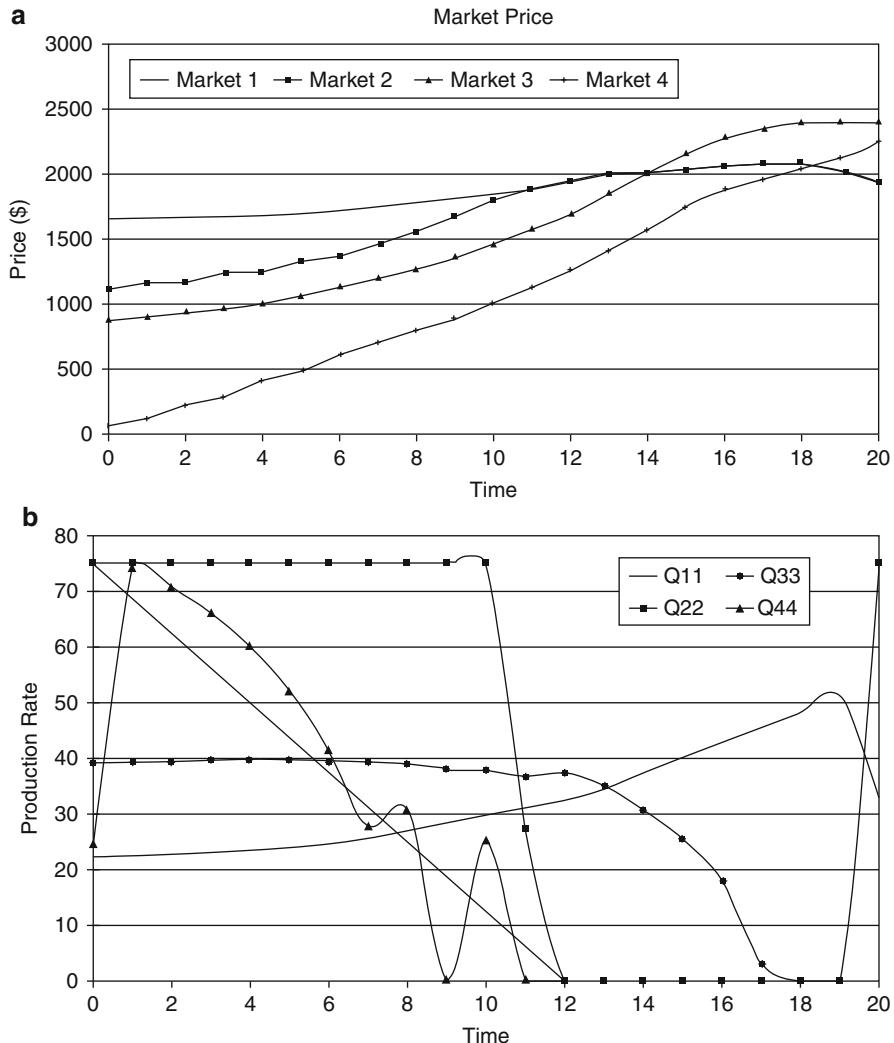


Fig. 13.3 Production rates and market price

Figure 13.3b shows that each firm follows a different production plan. Firm 1 slowly increases production until it nears the end of the planning horizon, where production begins to decline. Alternatively, firm 2 operates at its full capacity for the first ten periods, then abruptly halts production only returning to full production for the last period to meet the final inventory constraint. Figure 13.2a suggests that firms must be capable of dramatically altering production schedules if they are to compete in the final goods market successfully. The numerical results for dynamic network oligopoly show complicated temporal behavior that cannot be deduced prior to numerical analyses.

13.6 Conclusions

This chapter has discussed static and dynamic spatial oligopoly. In this market structure, there is a sufficiently small number of firms competing in such a way that the strategy of any individual firm has a significant effect on the strategies of the other firms. We began with the spatial oligopoly model and moved on to the static and dynamic models by considering both static and dynamic transportation networks with several spatially separated markets for a single commodity.

Since this chapter emphasized computation rather than theory, we have formulated the static model as a variational inequality (VI) which can be solved numerically by the diagonalization algorithm. Further, it has been shown that dynamic oligopolistic network competition is easily and naturally formulated as a differential variational inequality (DVI). A discrete-time approximation method has been proposed to solve the DVI problem. One advantage of time discretization is that we can completely eliminate state variables and obtain finite-dimensional control variables involving only upper and lower bound constraints. Another advantage of our formulation is that its solution methodology takes advantage of well-documented and available optimization techniques. In particular, one may employ a standard nonlinear mathematical programming package to obtain a solution to the dynamic network oligopoly model.

To extend the network models discussed in this chapter, one might consider introducing traffic congestion in the distribution network. Another natural extension for spatial oligopoly models would be considering stochasticity which might emerge from the uncertainty of parameters. The games associated with such a model are known as stochastic games, and one may obtain a probability distribution for the Nash equilibrium.

References

- Beckmann MJ, Puu T (1990) Spatial structures. Springer, Berlin/Heidelberg/New York
- Brander JA, Zhang A (1993) Dynamic oligopoly behavior in the airline industry. *Int J Ind Organ* 11(3):407–435
- Dafermos S, Nagurney A (1987) Oligopolistic and competitive behavior of spatially separated markets. *Reg Sci Urban Econ* 17(2):245–254
- Friedman JW (1979) Oligopoly and the theory of games. North-Holland, New York
- Friesz TL (2010) Dynamic optimization and differential games. Springer, New York
- Friesz TL (1993) A spatial computable general equilibrium model. In: Proceedings of the workshop on transportation and computable general equilibrium models, Venice, 19–21 May 1993
- Friesz TL, Rigdon MA, Mookherjee R (2006) Differential variational inequalities and shipper dynamic oligopolistic network competition. *Transp Res Pt B* 40(6):480–503
- Friesz TL, Harker PT (1984) Multicriteria spatial price equilibrium network design: theory and computational results. *Transp Res Pt B* 17(5):411–426
- Greenhut M, Lane WJ (1989) Theory of oligopolistic competition. *Manch Sch* 57(3):248–261
- Greenhut M, Norman G, Hung CS (1987) The economics of imperfect competition: a spatial approach. Cambridge University Press, Cambridge, UK

- Hakimi SL (1983) On locating new facilities in a competitive environment. *Eur J Oper Res* 12(1):29–35
- Harker PT (1986) Alternative models of spatial competition. *Oper Res* 34(13):410–425
- Harker PT (1984) Variational inequality approach for the determination of oligopolistic market equilibrium. *Math Program* 30(1):105–111
- Hashimoto H (1985) A spatial Nash equilibrium model. In: Harker PT (ed) *Spatial price equilibria: advances in theory, computation, and application*. Springer, Berlin/Heidelberg/New York, pp 20–41
- Henderson JM, Quandt RE (1980) *Microeconomic theory: a mathematical approach*, 3rd edn. McGraw-Hill, New York
- Matsushima N, Matsumura T (2003) Mixed oligopoly and spatial agglomeration. *Can J Econ* 36(1):62–87
- Markovich S (2008) Snowball: a dynamic oligopoly model with indirect network effects. *J Econ Dyn Control* 32(3):909–938
- Murphy FH, Sherali HD, Soyster AL (1982) A mathematical programming approach for deterministic oligopolistic market equilibriums. *Math Program* 24(1):92–106
- Nagurney A, Dong J, Zhang D (2002) A supply chain network equilibrium model. *Transp Res Pt E* 38(5):281–303
- Nagurney A (1999) *Network economics: a variational inequality approach*, Revised 2nd edn. Kluwer, Boston
- Nagurney A, Dupuis P, Zhang D (1994) A dynamical systems approach for network oligopolies and variational inequalities. *Ann Oper Res* 28(3):263–293
- Nash J (1950) Equilibrium points in n-person games. *Proceed Nat Aca Sci* 36(1):48–49
- Nash J (1951) Non-cooperative games. *Annal Math* 54(2):286–295
- Novshek W (1980) Equilibrium in Simple Spatial (or differentiated product) Models. *Journal of Economic Theory* 22(2):313–326
- Raa T (1984) The distribution approach to spatial economics. *J Reg Sci* 24(1):105–117
- Rovinskey RB, Shoemaker CA, Todd MJ (1980) Determining optimal use of resources among regional producers under differing levels of cooperation. *Oper Res* 28(4):859–866
- Tobin RT, Friesz TL (1983) Formulating and solving the spatial price equilibrium problem with transshipment in terms of arc variables. *J Reg Sci* 23(2):187–198
- Wie BW, Tobin LT (1997) A dynamic spatial Cournot-Nash equilibrium model and an algorithm. *Comput Econ* 10(1):15–45

Demand-Driven Theories and Models of Regional Growth

14

William Cochrane and Jacques Poot

Contents

14.1	Introduction	260
14.2	Taxonomy of Theoretical Perspectives on Long-Run Growth	261
14.3	Exogenous Demand Determined Regional Expansion	263
14.4	The Kaldor-Dixon-Thirlwall Model	265
14.5	The Benefits of High Wages and Public Expenditure for Growth	269
14.6	Institutional Theories of Regional Growth	272
14.7	Conclusions	274
	References	275

Abstract

In this chapter, we focus on theories and models of growth that have their origin in Keynesian economics. Their common features are that firstly, growth is largely export-driven; secondly, increasing returns yield path dependencies and possible divergence; thirdly, full resource utilization is not guaranteed; fourthly, economic expansion may face a balance of payments constraint, even at the regional level; and fifthly, institutions matter. We first briefly contrast demand-driven growth theories with neoclassical and other perspectives in taxonomy of growth theories. We then show how growth in exports yields regional income growth via a multiplier that is positively associated with the propensity to consume locally produced output and the propensity to invest but

W. Cochrane (✉)

School of Social Sciences, University of Waikato, Hamilton, New Zealand

e-mail: billpsc@gmail.com

J. Poot

National Institute of Demographic and Economic Analysis, University of Waikato, Hamilton, New Zealand

e-mail: jpoot@waikato.ac.nz

negatively related to regional tax rates and the extent to which government transfers are countercyclical. We show that Verdoorn's law – economic expansion generates productivity growth – leads to both sustained export growth and steady-state income growth, with the latter in balance of payments equilibrium equaling the rate of growth of exports divided by the income elasticity of the demand for imports. Next, theories are reviewed that suggest that policies that encourage regional growth in wages and public expenditure can be growth enhancing. Finally, we argue that the effectiveness of such demand-driven growth policies depends on institutional settings.

14.1 Introduction

The models of economic growth that were discussed extensively in the previous chapters of this section of the major reference work can be broadly labeled as supply-side-oriented general equilibrium explanations of regional economic development. Given resources of labor and capital, preferences, and technology, the allocation and accumulation of regional resources is in supply-side growth models determined by market forces that generate a general equilibrium in which prices, wages, and profits are simultaneously determined with employment, the allocation of capital, and production in all sectors, including new capital goods. The demand side of the economy is in these models fully determined by this general equilibrium with, for example, consumption determined by factor incomes, prices, and the propensity to consume. Productivity improvements are in such general equilibrium models generated by implicit or explicit innovation sectors and by human capital investments that enhance the stock of knowledge. Such models may be either stylized or calibrated by regional data in the form of spatial computable general equilibrium (SCGE) models (Donaghy 2008). Dynamic versions of such models can describe how the production of capital goods and population growth in each period contribute to an expansion of the regional production factors that, jointly with technologically driven changes in factor productivity, determine growth. These models can provide a complete and detailed bottom-up quantitative description of the growth path of the regional economy.

However, regional economies often exhibit features that are inconsistent with the smooth adjustments and economic expansion that such general equilibrium models imply. Firstly, many goods in the regional economy have prices determined outside the region, modified by transportation costs. Secondly, local incomes may depend greatly on external demand for local production and on government transfers. Thirdly, barriers to the mobility of labor may lead to considerable differences in regional unemployment rates and wages. Fourthly, the available capital may be underutilized, or alternatively, finance for new investment may be constrained by unfavorable local expectations. Finally, the diffusion of technological change may also vary between regions and depend on local investment and entrepreneurial activity. It is these features of regional economies that have led to the development of regional models in which demand for goods and services produced by the region

becomes the main driver of regional economic activity. Textbook examples are the economic base model and the regional Keynesian multiplier model. When investment in the local economy is determined by the income generated by local production, or expectations of future income, and technological change is embodied in such new investment and also depends on the scale of local production, regional growth becomes then primarily dependent on the evolution of regional demand or dependent on an exogenous demand shock that triggers an endogenous growth process. Demand-led growth models are a key focus of this chapter. Following a discussion of demand-led growth models and the implications of a regional balance of payments constraint, we also briefly discuss related Keynesian theories of growth and theories that emphasize the role of institutions.

This chapter is organized as follows. The next section briefly outlines the taxonomy of theoretical perspectives on long-run growth, while [Sect. 14.3](#) highlights the central role that regional exports and other forms of regional exogenous demand, such as local government expenditure and infrastructural investment, play in determining regional economic activity. [Section 14.4](#) discusses a model of export-driven growth via productivity improvements that are assumed to result from regional expansion. We also show how limited export opportunities may generate a balance of payments constraint on economic expansion. [Section 14.5](#) presents theories that suggest how high wages and government consumption may trigger productivity improvements that can generate regional growth. [Section 14.6](#) considers the role of institutions. [Section 14.7](#) concludes.

14.2 Taxonomy of Theoretical Perspectives on Long-Run Growth

It is important to note that the various theoretical perspectives available on long-run growth are not necessarily contradicting but may offer different but often complementing perspectives on a complex reality. This is illustrated in [Fig. 14.1](#) which charts different economic schools of thought in terms of some of the key paradigms that may be emphasized in the study of economic growth. The first paradigm is that of a permanent tendency of the economy to move toward a stable equilibrium, driven by market forces. An alternative paradigm is that of path dependency, evolutionary change, and complex dynamics. A third paradigm emphasizes the need for coordination and intervention in dynamic processes in order to reach desirable goals. These three paradigms may be juxtaposed with two additional guiding principles that are both dichotomies: a distinction between static and dynamic perspectives and a distinction between the policy goals of economic efficiency on the one hand and an equitable distribution of wealth and income on the other.

By combining the three paradigms with the two dichotomies, we obtain a classification that may be helpful to highlight the features and perspectives of different theories of economic growth. [Figure 14.1](#) provides a single example for each school of thought regarding economic growth. At the top, neoclassical theory

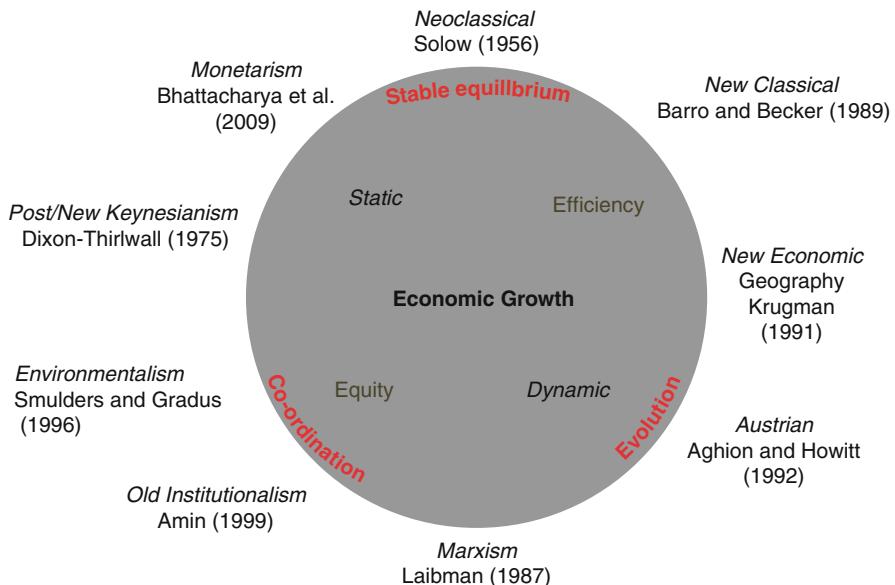


Fig. 14.1 Schools of thought on economic growth with selected examples

can be characterized by the emphasis on economies moving toward a long-run exogenously determined steady-state growth path, of which the Solow (1956) growth model is clearly the leading proponent. Toward the right, and emphasizing efficiency and micro-foundation-based rationality, even in terms of, for example, dynamic utility across generations, are new classical growth models such as Barro and Becker (1989). Moving away from stable equilibrium and toward dynamical systems modeling, one reaches core-periphery models of new economic geography, as first formulated by Krugman (1991). Next, toward even greater dynamical complexity and evolution, we reach Schumpeterian endogenous growth models of creative destruction (Aghion and Howitt 1992).

While having become largely a historical school of thought, completeness warrants inclusion of Marxist perspectives on growth, which are diagrammatically opposite those of neoclassical thinking (e.g., Laibman 1987). In the lower left quadrant of the figure in which equity and coordination are emphasized, we find institutionalism, also referred to as old institutional economics (e.g., Amin 1999) as discussed in Section 14.5, and models that are concerned with sustainability, externalities, and resource depletion (e.g., Smulders and Gradus 1996). Given the emphasis on rigidities, persisting unemployment but with nonetheless the possibility of long-run demand-driven steady-state growth, the post-Keynesian growth models are in the top left quadrant. The exemplar is the Dixon-Thirlwall model discussed in detail in this chapter. Finally, monetarist perspectives, originating from the ideas of Milton Friedman, also have been linked to long-run growth; see, for example, Bhattacharya et al. (2009).

14.3 Exogenous Demand Determined Regional Expansion

Demand-driven models of regional growth have their origin in the Keynesian theory of macroeconomic equilibrium. In this theory, equilibrium is determined by the equality of income generated by regional production and expenditure funded by this income, rather than by the available supplies of production factors. The latter are assumed abundantly available and short-run equilibrium may coincide with underutilization of the available capital, unemployment of labor, relatively rigid wages, and limited interregional migration. We develop the model largely along the lines of McCann (2001). Given the equality of regional income and expenditure,

$$Y_r = C_r + I_r + G_r + X_r - M_r \quad (14.1)$$

in which Y_r is value added or income, C_r private consumption, I_r gross fixed capital formation, G_r government consumption, X_r exports, and M_r imports, all referring to region r . Consumption expenditure is partially autonomous and partially a function of regional income:

$$C_r = AC_r + c_r Y_r (1 - \tau_r) \quad (14.2)$$

with c_r the marginal propensity to consume out of disposable income and AC_r representing autonomous consumption that is unrelated to regional income which, for example, could be consumption funded by national pensions. The parameter τ_r represents the regional average tax rate. Investment has similarly an autonomous component AI_r , (e.g., reflecting expectations and replacement investment) and a component that is linked to regional disposable income:

$$I_r = AI_r + i_r Y_r (1 - \tau_r) \quad (14.3)$$

in which i_r is the marginal propensity to invest in new capital goods for given disposable income. Government consumption at the regional level will be the combination of expenditure initiated by local government and by national government. It is not implausible to assume that it negatively related to regional disposable income, as governments often have a nationally determined equity objective in which declines in regional disposable income due to depressed regional economic activity are partially offset by additional public consumption or by publicly funded private consumption through benefit payments. Hence,

$$G_r = AG_r - g Y_r (1 - \tau_r) \quad (14.4)$$

in which AG_r stands for autonomous public consumption and g measures the propensity of government consumption to be larger in regions with lower disposable income. Exports are partially autonomous (AX_r) and partially a function of income in the rest of the country and the world Z , the local price level p_r , the foreign

price level p_f , and the exchange rate e . Following Dixon and Thirlwall (1975) and others, this relationship is defined as one with constant elasticities rather than linearly. Hence,

$$X_r = AX_r Z^{\varepsilon_r} [p_f/(ep_r)]^{\eta_r} \quad (14.5)$$

The term $[p_f/(ep_r)]$ represents the price competitiveness of the regional economy in trade, with η_r the competitiveness elasticity of exports. Similarly, ε_r represents the world income elasticity of local exports (both elasticities may vary across regions).

Finally, imports are partially autonomous (AM_r) and partially a function of regional disposable income and price competitiveness of the regional economy in trade, again with potentially spatially varying elasticities, here π_r and μ_r , respectively:

$$M_r = AM_r [Y_r(1 - \tau_r)]^{\pi_r} [(ep_r)/p_f]^{\mu_r} \quad (14.6)$$

In order to make the model easily solvable, we assume that the elasticity of imports with respect to regional disposable income is one, that is, $\pi_r = 1$. This implies that for a given competitiveness of the regional economy, imports are proportional to disposable income. Defining $m_r \equiv AM_r [(ep_r)/p_f]^{\mu_r}$, Eq. (14.6) then simply becomes

$$M_r = m_r [Y_r(1 - \tau_r)] \quad (14.7)$$

Similarly, defining $x_r \equiv AX_r [p_f/(ep_r)]^{\eta_r}$, gives

$$X_r = x_r Z^{\varepsilon_r} \quad (14.8)$$

Substituting Eqs. (14.2)–(14.4), (14.7), and (14.8) into Eq. (14.1), and solving for Y_r , yields the equilibrium level of regional income:

$$Y_r = \{1 - [(c_r - m_r) + (i_r - g)](1 - \tau_r)\}^{-1} \{AC_r + AI_r + AG_r + x_r Z^{\varepsilon_r}\} \quad (14.9)$$

The first term on the right-hand side is the Keynesian multiplier. This multiplier, which we will denote by k_r , is expected to be greater than one because firstly, the marginal propensity to consume locally produced goods ($c_r - m_r$) is expected to be less than one; secondly, the regional marginal propensity to invest, corrected for countercyclical government transfers, is also expected to be between zero and one; and thirdly, the average tax rate is less than one. The Keynesian multiplier will increase when the marginal propensity to consume locally produced goods increases, when the demand for imports from other regions and countries declines, when the regional marginal propensity to invest increases, when government transfers are less countercyclical, and when the average tax rate declines.

The greatest source of regional fluctuation in autonomous demand is likely to be regional exports, which are – for given competitiveness – fully driven by incomes in

other regions and countries. If we assume that the other components of autonomous demand are constant, a change in regional exports (ΔZ^{e_r}) leads to the following change in regional income:

$$\Delta Y_r = k_r x_r \Delta Z^{e_r} \quad (14.10)$$

If q_r refers to average labor productivity in the regional economy, q_{xr} refers to average productivity in the export sector, and these productivities are assumed to be constant in the short run, Eq. (14.10) can be transformed to one describing changes in employment:

$$\Delta T_r = k_r (q_{xr} / q_r) \Delta B_r \quad (14.11)$$

in which T_r refers to total regional employment and B_r refers to employment in the export sector. The latter is often referred to as basic sector employment. All other sectors are then referred to as nonbasic sectors. Equation (14.11) shows that there is a multiplier relationship between basic sector employment and total regional employment. This multiplier is referred to as the economic base multiplier. The multiplier will be higher when the basic (export) sector has a relatively large demand for local inputs from nonbasic sectors. Given that labor productivity in the export/basic sector is likely to be higher than in the regional economy generally, that is, $q_{xr}/q_r > 1$, the economic base multiplier would be larger than the Keynesian multiplier.

However, both models describe the short-run response of the regional economy to regional demand shocks but not long-run growth. In the model summarized in Eq. (14.9), sustained export-led growth requires either a sustained increase in external (“world”) demand for regional exports or continual productivity improvements in the regional economy that lead to sustained increases in regional competitiveness (i.e., increases in x_r and decreases in m_r). The next section describes such a model of trade and productivity growth.

14.4 The Kaldor-Dixon-Thirlwall Model

This model was first formulated in Dixon and Thirlwall (1975) but was based on ideas discussed by Kaldor (1970). In turn, the fundamental mechanism that drives regional growth along Kaldorian lines is the so-called Verdoorn’s law (following Verdoorn 1949) which states that labor productivity growth is positively correlated with output growth. Following Armstrong and Taylor (2000), we will assume that this correlation can be interpreted as a causal relationship in which output growth in any period causes productivity growth in the following period. Adoption of the notation that $*$ refers to the growth rate of a variable, for example, $*Y_{rt} = (Y_{rt} - Y_{rt-1})/Y_{rt-1}$, Verdoorn’s law can then be expressed as follows:

$$*q_{rt} = *{\alpha}_r + {\lambda}_r *Y_{rt-1} \quad (14.12)$$

with q_{rt} again being regional labor productivity and Y_{rt-1} regional output (which equals income), as in the previous section. Equation (14.12) states that the growth in regional labor productivity ($*q_{rt}$) in any given period is partially exogenous ($*\alpha_r$) and partially a function of output growth in the previous period ($*Y_{rt-1}$), with coefficient λ_r . Clearly, an empirically detected correlation between output growth and productivity growth is in itself not informative of the channels through which an increase in the scale of production leads to greater labor productivity. A range of models with increasing returns or agglomeration economies discussed elsewhere in this section of the major reference work and also included in the bottom right quadrant of Fig. 14.1 may yield such a relationship. What distinguishes the present model from those supply-side-driven models is that in the present context, the accumulation of resources is not made explicit. Nonetheless, it is clear that steady-state growth triggered by an expansion of demand can only exist if resource capacity utilization is stable in the long run (e.g., Setterfield 2010).

As is typical in a model of Keynesian origin, prices are a markup on costs. Consequently, regional price inflation $*p_{rt}$ is directly related to labor cost inflation $*w_r$, which is assumed exogenous, and to productivity growth:

$$*p_{rt} = *w_r - *q_{rt} \quad (14.13)$$

To close the model, current output growth $*Y_{rt}$ is assumed to be driven by export growth. In the simplest form,

$$*Y_{rt} = \gamma_r *X_{rt} \quad (14.14)$$

in which γ_r measures the export elasticity of regional output. To solve the model, we first linearize Eq. (14.5) by expressing the export demand function in terms of growth rates:

$$*X_{rt} = *AX_{rt} + \varepsilon_r *Z_t + \eta_r (*p_{ft} - *e_t - *w_{rt} + *q_{rt}) \quad (14.15)$$

The Kaldor-Dixon-Thirlwall model now consists of the four Eqs. (14.12)–(14.15) that can be solved to yield the following first-order linear difference equation for the growth rate of regional income:

$$*Y_{rt} = \gamma_r [*AX_{rt} + \varepsilon_r *Z_t + \eta_r (*p_{ft} - *e_t - *w_{rt} + *q_{rt})] + \gamma_r \eta_r \lambda_r *Y_{rt-1} \quad (14.16)$$

It is clear from Eq. (14.16) that when the exogenous variables are assumed time invariant, the time path of the regional growth rate is fully determined by the product of three parameters, namely, the elasticity of regional output with respect to “world income” (γ_r), the elasticity of regional exports with respect to regional competitiveness (η_r), and the extent of increasing returns in the regional economy (λ_r). A steady-state growth rate requires $\gamma_r \eta_r \lambda_r < 1$. Given

that regional exports are roughly a constant proportion of regional output, $\gamma_r = 1$. Similarly, in a small regional economy that is a price taker in global markets, the export price elasticity may be at most around one as well. These empirically based assumptions can be combined with estimates of the Verdoorn coefficient that are around 0.5 (e.g., Fingleton and McCombie 1998). Together, this suggests that Eq. (14.16) indeed may converge to a steady-state growth rate:

$${}^*Y_r = \left\{ \gamma_r [{}^*AX_r + \varepsilon_r {}^*Z + \eta_r ({}^*p_f - {}^*e - {}^*w_r + {}^*\alpha_r)] \right\} / [1 - \gamma_r \eta_r \lambda_r] \quad (14.17)$$

which shows that the steady-state growth rate will be (i) positively related to local returns to scale or agglomeration benefits (a larger λ_r and/or a larger ${}^*\alpha_r$), (ii) positively related to sustained growth in external demand (a larger *Z), (iii) positively related to export-facilitating policies such as growth in port or airport infrastructure or structural change that benefits export industries (a larger *AX_r and/or ε_r), and (iv) negatively related to regional cost inflation (a larger *w_r or smaller η_r). Note also that according to Eq. (14.17), a change in world prices and/or the nominal exchange rate will have regionally specific impacts that depend on the regionally specific parameters γ_r , η_r , and λ_r .

If we combine the growth path of regional income as defined by Eq. (14.16) with regional expenditure given in Eq. (14.1), which is largely endogenous, it becomes clear that any net surplus (deficit) in a region's trade in goods and services with other regions and countries, given by $(X_r - M_r)$ in Eq. (14.1), is balanced by the region's net acquisition (disposal) of assets from elsewhere, or

$$Y_r - A_r = X_r - M_r \quad (14.18)$$

in which $A_r = C_r + I_r + G_r$ refers to regional absorption of income. It has been argued by Thirlwall (1980, 1997) that such a spatial redistribution of equity and debt cannot last forever. If a region has a balance of payments deficit, this deficit must be financed by borrowing from other regions or countries, which effectively increases the claims of outsiders on regional assets. Conversely, a region with a balance of payments surplus will acquire assets outside the region. Although this may be considered a less severe situation than a balance of payments deficit, sustained surpluses constrain the disposable income of other regions and countries, net of borrowing costs. Consequently, in the long run, the regional balance of payments will tend to equilibrium, which implies that export receipts (in foreign currency) equal import payments, that is,

$${}^*e p_r X_r = {}^*p_f M_r \quad (14.19)$$

Rewriting Eq. (14.19) in growth terms gives

$${}^*e + {}^*p_r + {}^*X_r = {}^*p_f + {}^*M_r \quad (14.20)$$

If we assume that in the long run the exchange rate will adjust to a difference in the growth of local prices and foreign prices, that is, purchasing power parity holds in the long run or $*e + *p_r = *p_f$, then

$$*X_r = *M_r \quad (14.21)$$

From Eq. (14.6) we can derive that $*M_r = \pi_r *Y_r$, and from Eq. (14.5) that $*X_r = \varepsilon_r *Z$. Substituting this in Eq. (14.21), we see that long-run balance of payments equilibrium is only compatible with the following long-run growth rate of regional income:

$$*Y_r = \varepsilon_r *Z / \pi_r \quad (14.22)$$

In long-run balance of payments equilibrium, the growth rate of regional income is proportional to the growth in external income. The proportionality constant is regionally specific and equal to the ratio of elasticity of regional exports with respect to global income over the regional income elasticity of the demand for imports.

Given that $*Z$ does not vary across regions, Eq. (14.22) is an empirically testable proposition that long-run balance-of-payments-constrained regional growth rates are proportional to ε_r/π_r . Unfortunately due to data deficiencies, it is not easy to test this relationship, referred to as Thirlwall's law in the literature, at the regional level. The relationship has instead been tested empirically with cross-country and panel data. These econometric tests have yielded mixed results. Nonetheless, McCombie (2011) concludes in a recent review of this literature that "After over thirty years since its development, Thirlwall's Law is still proving a powerful explanation of why growth rates differ" (pp. 388–389). Both the evidence of contagion in the international economy (financial crises that start in one country quickly spread to countries that have a strong trading relationship with the former country) and the evidence of spatial spillovers in regional modeling indeed provide convincing support for the idea that economic growth in small open economies is strongly dependent on demand elsewhere. The challenge is, however, to derive policy recommendations from this empirical regularity. In this respect, Krugman (1989) argues that causation in Eq. (14.22) runs in the opposite direction: output growth leads to productivity growth. Sustained output growth coincides with an increase in varieties of output, quality improvement, other forms of innovation, and/or agglomeration. These processes lower unit production costs and enable regions to compete more effectively. Such scale effects may also lessen the reliance on imports. Together, these phenomena lead to a greater ε_r/π_r ratio.

In conclusion, demand-driven growth models and endogenous growth models have more in common than proponents of either tend to admit. The distinction is a matter of emphasis, both in terms of the underlying engine of growth and in terms of the time frame (e.g., Setterfield 2010). Verdoorn's (1949) law and Myrdal's (1957) theory of cumulative causation are consequences of behavioral and

technological phenomena that lead to increasing returns and that have been analytically formulated and empirically tested in endogenous growth models. Moreover, demand-driven growth theories emphasize short- to medium-term constraints in demand that lead to less than full resource capacity utilization, combined with slowly changing regional structures and some price and wage stickiness. In contrast, neoclassical growth theories emphasize the mobility of production factors and relative price adjustments in the long run that lead to trade following the emerging patterns of comparative advantage. As already illustrated in Fig. 14.1, these different perspectives provide complementing rather than contradicting insights into regional economic development.

14.5 The Benefits of High Wages and Public Expenditure for Growth

From the end of the Second World War until the early 1980s, regional development in the Western world was heavily influenced by the principles of Keynesian economic management. Development policy, then, was highly interventionist and aimed to promote growth in less-favored regions through income redistribution and welfare programs (bolstering local aggregate demand), the provision of state-funded incentives to induce firms to locate in these regions, and through the funding of large-scale infrastructural investment with the dual aims of improving the productivity of existing firms and inducing new firms to locate in the region (Amin 1999, p. 365). As a consequence of such capital-oriented bolstering of demand, regional output and employment may be expected to grow, as shown in Sect. 14.3. In the presence of increasing returns to scale, as modeled by Verdoorn's law in the previous section, the region's long-run growth rate may then be expected to increase. Martin (2005, pp. 2–7) provides a useful summary of the Keynesian approach (see Table 14.1).

In keeping with the reorientation of mainstream economics to micro-foundations and supply-side-oriented general equilibrium theory, Keynesian approaches to regional development, and demand-driven policies in general, fell from favor. However, this was more pronounced in the theorizing of regional development than in the practice, with practitioners and regional policymakers often resisting attempts by central government at introducing regional policy inspired by neoclassical economics. The move away from Keynesian approaches was not solely a function of hegemony of neoclassicism but also reflected the rather modest results achieved by Keynesian policy in lifting the long-run growth rate of particularly peripheral or “rust belt” regions. While there were conspicuous successes in implementing these policies, such as the Tennessee Valley Authority, there were (at least) an equal number of spectacular failures. Furthermore, there was a lack of theorizing of how growth could be maintained, a tendency for regional development agencies to fund activities that would have taken place at any rate (Pike et al. 2006, p. 77) or, worse, investment in activities for primarily short-run political

Table 14.1 Key elements of Keynesian theory

Key assumptions	Key driving factors
<ul style="list-style-type: none"> • Price adjustments might be slow, leading to adjustments in quantity • Markets are not necessarily in equilibrium • Shortages on demand or supply side • Possibility of “false” trading (i.e., with non-equilibrium prices) • Capital and labour are complementary 	<ul style="list-style-type: none"> • Capital intensity • Investment • Government spending, such as investment in the public domain and subsidies/tax cuts for enterprises
Implications for (regional) competitiveness	
<ul style="list-style-type: none"> • Governments can intervene successfully in the cycles of the economy: timing is crucial • Assumption of imperfect markets allows for regional differences • Convergence of regions can only be achieved through economic policy 	

Source: Martin (2005)

purposes. In addition, much investment in Keynesian policies was directed at maintaining declining industries in less-favored regions (LFR) that stood little chance of success once support was removed rather than in facilitating the growth of new industry or the mobility of workers both between regions and between industries or occupations. More generally it was difficult to see how the indiscriminate application of Keynesian policy would, in isolation, lead to higher levels of employment for the most disadvantaged or how these policies would contain inflationary dynamics (Mitchell and Juniper 2006, p. 14).

The regional development policy arising from neoclassical theorizing, however, fared arguably no better than the Keynesian in changing the fate of LFR with some, such as Amin (1999, p. 365) arguing that the results of the implementation of these policies were a “far worse outcome, by removing financial and income transfers which have proven to be vital for social survival, by exposing the weak economic base of the LFRs to the chill wind of ever enlarging free market zones, and by failing singularly to reverse the flow of all factor inputs away from the LFRs (i.e., no proof of price-seeking inflow of opportunities leading to regional specialisation in the appropriate industries).”

This failure, and subsequent events, such as the global financial crisis beginning in 2007, revived interest, particularly among those with social democratic or corporatist orientations, to look for alternate approaches. One such approach argues for demand- or wage-led strategies to underpin long-run sustainable growth as opposed to profit-led strategies that purportedly dominate “New Right” policy agendas.

The history of this wage-led growth approach is long with a lineage stretching back to debates in eighteenth- and nineteenth-century political economy on the possibility of what was then called “underconsumption.” Contemporary debates tend to take as their starting point the seminal works of Rowthorn (1981) and Dutt (1987), all influenced by Kalecki and to a lesser extent Keynes, and have been taken up by more recent authors such as Stockhammer, Palley, Lavoie, Naastepad, and others (see the *International Journal of Labour Research*, 2011, 3(2) for a recent compilation of work on this topic).

Taking Palley (2011) as an example of the wage-led approach to economic growth, it can be argued that prior to the 1980s, the economies of the Western world could largely be described by a Keynesian growth model “built on full employment and wage growth tied to productivity growth” (Palley 2011, p. 222). The logic of this model was that productivity growth drove wage growth which led to increased demand and hence full employment. This in turn incentivized investment which underpinned further increases in productivity thus the elements of this mode of development were linked together in a “virtuous circle,” which is consistent with Myrdal’s (1957) theory of cumulative causation and with the Verdoorn (1949) effect. Post 1980, this pattern of development was displaced by supply-oriented models of growth which renounced two key elements of the Keynesian growth model – namely, the commitment to full employment and the linkage of wage growth to productivity growth – replacing them with a focus on maintaining low levels of inflation, removing impediments to capital mobility, and reducing “rigidities” in the labor market. The shortfall in aggregate demand generated by stagnant or falling wages was met by increasing levels of debt and speculation-driven asset price inflation which ultimately compromised the systemic stability and reproduction of the global economy.

Hence, in Palley’s (2011) narrative, the current economic crisis is primarily a crisis of demand. Supply-side measures are therefore unlikely to improve matters and – in so far as they seek to reduce real wages in an attempt to restore profitability – may very well worsen matters considerably. Thus, the only alternative in his view is to rebuild the virtuous circle of the Keynesian growth model. To achieve this, he has a number of clear policy suggestions:

- Rebuild the wage-productivity link.
- Large-scale use of public expenditure to compensate for inadequate levels of private demand – particularly in private infrastructure.
- Refocus monetary policy on full employment.
- Reregulate the finance sector and focus it on the needs of the real sector rather than on speculation.
- Reform corporate governance to emphasize long-term objectives and to direct investment to the real sector.
- Direct any tax cuts at those with high propensities to consume, thus bolstering aggregate demand.
- Coordinate economic policy at an international level to eliminate long-run balance of payments deficits and surpluses and to protect the ability of nation states to engage in demand-led policies.

What then does this mean for regional development? A return to Keynes in terms of regional development requires a high level of national, probably international, coordination as without this it is hard to see that regions “going it alone” and pursuing Keynesian policies would not suffer from the equivalent of a balance of payments constraint or significant problems with debt financing, akin to nations pursuing such strategies in a currency union. Thus, demand-led growth is not a regional development policy as such but rather a national or international project within which regional strategies can be articulated.

Perhaps of more direct relevance to regional development is the work of Mitchell and Juniper (2006) and others advocating a new “spatial Keynesianism.” Speaking from a modern monetary theory (MMT) perspective, they eschew both the supply-side policies of neoclassicism and the generalized expansionist policies favored by Keynesians. Instead, they advocate policies that aim to achieve full employment, price stability, and environmental sustainability while preserving social settlements. The preservation of social settlements is justified both on equity grounds and to preserve networks and spatial spillovers (Mitchell and Juniper 2006, p. 20). Central to this approach is a job guarantee scheme designed to ensure full employment through the use of regionally targeted public sector employment at the minimum wage.

Under the job guarantee scheme, the state functions as an employer of last resort, providing a buffer stock of jobs that are available upon demand and as of right. This approach minimizes the impact of fluctuation in private sector demand – expanding in downturns and contracting as the private sector recovers. As the employment is at the minimum wage rate, the effects of the scheme are neither likely to distort the structure of the wage distribution nor prove inflationary. This last point is particularly important as the Keynesian approach lacks an explicit means of containing wage and price inflation, relying usually on income policy which has proved relatively ineffective and often counterproductive as it weakens the resolve of employers to rationalize and tends to hurt efficiency and investment projects.

14.6 Institutional Theories of Regional Growth

Institutional views on regional economic growth can be seen as belonging to one of two distinctly different viewpoints – “old institutional economics (OIE)” and “new institutional economics (NIE),” both of which emphasize the centrality of institutions in determining the economic trajectory of regions. While both OIE and NIE share a focus on institutions, they differ markedly in their conceptualizations of what constitutes an institution and are positioned very differently with respect to neoclassical economics. NIE is essentially a branch of neoclassical theory that emphasizes micro-foundations, although it rejects two central ideas of conventional neoclassical theory, namely, costless transactions and neutral institutions (Lakshmanan and Button 2009). OIE is however much closer to a wholesale rejection of the neoclassical paradigms.

Given that the focus of this chapter is on demand-driven theories of regional growth, this section will largely be concerned with the OIE though it should be noted that while the OIE and NIE share little in the way of common ground, at an abstract level their policy recommendations are often surprisingly similar, stemming from a shared view that regional-level industrial configurations, supply-side characteristics, and institutional arrangements, can play a critical role in securing the economic success of a region (Amin 1999, p. 368). As such the policy recommendations of institutionalists tend to go considerably beyond the traditional preserves of economic development, concentrating their attention on building the

wealth of a region, rather than the individual firm and often arguing for comprehensive renovation of the economic and social infrastructure of a region as a necessary precursor to the revitalization of economic activity in an area (Amin 1999, p. 370). Lakshmanan and Button (2009) provide an NIE perspective on regional development in which they emphasize the importance of institutions to benefit from positive externalities from education, healthcare, and infrastructure, as well as the importance of institutions for fostering entrepreneurship and innovation.

Following a line of theory development dating back to the works of Veblen, Ayres, Commons, and Mitchell, the OIE tradition argues that institutions not only act as constraints on economic actors but also are constitutive, in large part, of an individual's habits, preferences, and values as well as the range of actions available to them. Though the individual is born into a preexisting web of social relations, the relationship between the individual and the ensemble of institutions they are born into is not one of pure subordination but is reflexive in that the individual and institution(s) are mutually constitutive of each other, though institutions tend to have temporal priority over the individual. This line of reasoning is rooted in a conception of institutions that is considerably broader than those that conceive of them as largely being made up of the various organs of the state apparatus plus various legally constituted entities such as firms and trade unions. The OIE view of institutions is founded in Veblen's (1961) formulation, and its subsequent extensions, that institutions are "the settled habits of thought of the generality of men" or "systems of established and prevalent social rules that structure social interactions."

In terms of its methodological commitments, the OIE views with considerable antipathy the assumptions of neoclassical economics or of the new economic geography as developed by Krugman (1991, 1994) and others. OIE criticizes these approaches on the grounds of their (alleged) lack of realism and failure to adequately appreciate that economic processes are "embedded," operating within a social framework shaped by the history and culture of a region. Against the methodological individualism of neoclassicism, the OIE adopt holistic qualitative approaches such as in-depth case study research and largely reject formal econometric modeling on the grounds that it is incapable of capturing place-specific qualitative factors (such as culture and institutions) which they view as central to understanding differences in the economic trajectory of regions.

How then do these largely theoretical concerns translate into an understanding of the growth trajectory of a particular region? Martin (2005, p. 79) suggests that the starting point for enquiry within the OIE tradition is to pose the question "to what extent and in what ways are the processes of geographically uneven capitalist economic development shaped and mediated by the institutional structures?" The answer in the OIE literature emphasizes that regional economies are not collections of atomistic optimizing firms and markets, propelled by rational preferences operating under a common set of rules (Amin 1999, p. 367). Instead, OIE stresses the importance of geographical proximity, networking, embeddedness, and the development trajectories of regions, institutions, and technology in determining regional economic outcomes. Specifically Amin (1999, p. 368) derives a number of general axioms of economic governance associated with an institutionalist approach:

- “First, a preference for policy actions designed to strengthen networks of association, instead of actions which focus on individual actors.
- Second, that policy action should involve a plurality of decentralised and autonomous organisations since effective economic governance extends beyond the reach of both the state and market institutions (Hirst 1994).
- Third, within a frame of plural and autonomous governance, the role of the state, as the prime collective organisation with societal reach and legal power, should be that of providing resources, arbitrating between decentralised authorities, securing collective results, and, above all, establishing the strategic goal, rather than that of central planner or market facilitator (Hausner 1995).
- Fourth, the aim of policy action should be to encourage voice and negotiation, together with procedural and recursive rationalities of behaviour, rather than self-serving or rule-following behaviour, in order to secure strategic vision, learning and adaptation (Amin and Hausner 1997).
- Fifth, solutions have to be context-specific and sensitive to local path-dependencies.
- Sixth, there is a need to encourage intermediate forms of governance, building up to a local ‘institutional thickness’ (Amin and Thrift 1994) which includes enterprise support systems, political institutions, and social citizenship.
- Finally, and as a consequence, building economic success is as much a matter of devising appropriate economic policies as wider social and political reforms to encourage the formation of social capabilities for autonomous action (Putnam 1993). ”

Hence, institutionalist strategies to promote regional economic development are highly place specific as they are largely reliant on the ability to mobilize, foster, and coordinate local institutions which are the product of a unique developmental trajectory. The local cannot however be abstracted from its macro context. Regions, while perhaps possessing considerable latitude at times, are subsumed within nation states who set national-level macroeconomic and social policy which may not be congruent with the development of optimal institutional arrangements at a regional level.

14.7 Conclusions

In this chapter, we contrasted the neoclassical perspectives on economic growth with a range of alternative perspectives that we broadly referred to as demand-driven models of regional economic growth. Such demand-driven models have a number of important elements in common that were emphasized in the chapter (see also Setterfield 2010). Firstly, we showed how aggregate demand and in particular exports can affect the long-run growth path. Moreover, changing demand conditions can create path dependencies in regional development trajectories. Secondly, technological change is driven by a complex and broad range of processes that from this perspective are best modeled by the increasing returns to a greater scale of production, as formulated by Verdoorn’s law. Thirdly, full

resource utilization is not guaranteed given limited substitution possibilities, particularly in the short run. Fourthly, long-run stability is not ensured but dependent on a range of behavioral parameters. Additionally, economic expansion may face a balance of payments constraint and economies may diverge, even at the regional level. Fifthly, institutions matter and they embody a concern for equity and the potential feedback mechanisms of income inequality on growth.

We showed in Sect. 14.2 how growth in exports yields regional income growth via a multiplier that is positively associated with the propensity to consume locally produced output and the propensity to invest but negatively related to regional tax rates and the extent to which government transfers are countercyclical. We then formulated the Dixon-Thirlwall model in which Verdoorn's law – economic expansion generates productivity growth – leads to both sustained export growth and steady-state income growth, with the latter in balance of payments equilibrium equaling the rate of growth of exports divided by the income elasticity of the demand for imports. We also reviewed theories that suggest that policies that encourage regional growth in wages and public expenditure can be growth enhancing. Finally, we argued that the effectiveness of such demand-driven growth policies depends on institutional settings.

All theories of regional development have positive and normative elements. They may lead to testable hypotheses that highlight incompatibilities to the extent that acceptance of empirical evidence of a key feature of one theory is incompatible with the predictions of another. However, the greater the level of abstraction and the more aggregate the nature of the data, the greater the likelihood that various competing theories may lead to predictions that are observationally equivalent. In general, as Fig. 14.1 attempted to point out, the broad range of available theories may all contribute to a better understanding of the complex reality that we refer to as regional economic growth.

References

- Aghion P, Howitt P (1992) A model of growth through creative destruction. *Econometrica* 60(2):132–148
- Amin A (1999) An institutionalist perspective on regional economic development. *Int J Urban Reg Res* 23(2):365–378
- Amin A, Hausner J (eds) (1997) Beyond market and hierarchy: interactive governance and social complexity. Edward Elgar, Aldershot
- Amin A, Thrift N (1994) Institutional issues for the European regions: from markets and plans to socioeconomics and powers of association. *Econ Soc* 24(1):41–66
- Armstrong H, Taylor J (2000) Regional economics and policy, 3rd edn. Blackwell Publishing, Malden
- Barro RJ, Becker GS (1989) Fertility choice in a model of economic growth. *Econometrica* 57(2):481–501
- Bhattacharya J, Haslag J, Martin A (2009) Optimal monetary policy and economic growth. *Eur Econ Rev* 53(2):210–221
- Dixon R, Thirlwall AP (1975) A model of regional growth-rate differences on Kaldorian lines. *Oxf Econ Pap* 27(2):201–214

- Donaghy K (2008) CGE modeling in space: a survey. In: Capello R, Nijkamp P (eds) *Handbook of regional growth and development theories*. Edward Elgar, Cheltenham, pp 389–422
- Dutt AK (1987) Alternative closures again: a comment on ‘Growth, distribution and inflation’. *Camb J Econ* 11(1):75–82
- Fingleton B, McCombie JSL (1998) Increasing returns and economic growth: some evidence for manufacturing from the European Union regions. *Oxf Econ Pap* 50(1):89–105
- Hausner J (1995) Imperative vs. interactive strategy of systematic change in Central and Eastern Europe. *Rev Int Polit Econ* 2(2):249–266
- Hirst P (1994) *Associative democracy*. Polity Press, Cambridge, UK
- Kaldor N (1970) The case for regional policies. *Scott J Polit Econ* 17(3):337–348
- Krugman P (1989) Differences in income elasticities and trends in real exchange rates. *Eur Econ Rev* 33(5):1001–1046
- Krugman P (1991) Increasing returns and economic geography. *J Polit Econ* 99(3):483–499
- Krugman P (1994) Complex landscapes in economic geography. *Am Econ Rev (Pap Proc)* 84(2):412–416
- Laibman D (1987) Growth, technical change, and cycles: simulation models in Marxist economic theory. *Sci Soc* 51(4):414–438
- Lakshmanan TR, Button KJ (2009) Institutions and regional development. In: Capello R, Nijkamp P (eds) *Handbook of regional growth and development theories*. Edward Elgar, Cheltenham, pp 443–460
- Martin R (2005) Institutional approaches in economic geography. In: Sheppard E, Barnes TJ (eds) *A companion to economic geography*. Blackwell Publishing, Malden, pp 77–94
- McCann P (2001) *Urban and regional economics*. Oxford University Press, Oxford
- McCombie JSL (2011) Criticisms and defences of the balance-of-payments constrained growth model: some old, some new. *PSL Q Rev* 64(259):353–392
- Mitchell W, Juniper J (2006) Towards a spatial Keynesian macroeconomics. Centre of Full Employment and Equity, University of Newcastle, Newcastle
- Myrdal G (1957) *Economic theory and underdeveloped regions*. Duckworth, London
- Palley T (2011) The economics of wage-led recovery: analysis and policy recommendations. *Int J Labour Res* 3(2):219–244
- Pike A, Rodríguez-Pose A, Tomaney J (2006) *Local and regional development*. Oxon, Routledge
- Putnam R (1993) *Making democracy work: civic traditions in modern Italy*. Princeton University Press, Princeton
- Rowthorn R (1981) Demand, real wages and economic growth. *Thames Pap Polit Econ Autumn*:1–39
- Setterfield M (2010) An introduction to alternative theories of economic growth. In: Setterfield M (ed) *Handbook of alternative theories of economic growth*. Edward Elgar, Cheltenham, pp 1–14
- Smulders S, Gradus R (1996) Pollution abatement and long-term growth. *Eur J Polit Econ* 12(3):505–532
- Solow RM (1956) A contribution to the theory of economic growth. *Q J Econ* 70(1):65–94
- Thirlwall AP (1980) Regional problems are “balance of payments” problems. *Reg Stud* 14(5):419–425
- Thirlwall AP (1997) Reflections on the concept of balance-of-payments-constrained growth. *J Post-Keynes Econ* 19(3):377–385
- Veblen T (1961) *The place of science in modern civilisation and other essays*. Russell & Russell, New York
- Verdoorn PJ (1949) Fattori che regolano la sviluppo della produttività del lavoro. *L'Ind* 1(1):3–10

The Measurement of Regional Growth and Wellbeing

15

Philip S. Morrison

Contents

15.1	Introduction	278
15.2	Well-Being	279
15.3	Economic Growth and Well-Being	280
15.4	Subjective Well-Being and the Region	281
15.5	Well-Being and Relativities	285
15.6	Conclusions	287
	References	288

Abstract

Our understanding of people's well-being was, until very recently, inferred from observable objective indicators such as their income and education. These measures were then aggregated to generate an average that characterized the city or region. With the growing availability of sample survey data, we now have at our disposal an increasing range of subjective measures of well-being that capture quality of life assessments made by individuals themselves. It is these internal measures of subjective well-being from microdata that are now being widely used throughout the social sciences to study what we call well-being or "happiness."

Contemporary interest in subjective measures of well-being stems from a wish to supplement market-based criteria such as GDP per capita with other more direct measures of societal well-being. Subjective measures are particularly useful in areas where the distribution of outcomes is not easily identified using other, especially market, criteria. The effect of investment in public infrastructure or the provision of green space or in fostering community networks or in

P.S. Morrison

School of Geography, Environment and Earth Sciences, Victoria University of Wellington,
Wellington, New Zealand
e-mail: philip.morrison@vuw.ac.nz

redeveloping neighborhoods can be captured in responses to questions on well-being, preferably over time. These subjective measures, which have been shown to be highly correlated with clinical and other assessments of well-being, are likely to be of particular interest in regional science because of the way changes to places result from, or generate, a range of positive or negative externalities.

The decisions we as individual citizens make depend on what we measure, how good our measurements are and how well our measures are understood. . . . For example, traffic jams may increase GDP as a result of the increased use of gasoline, but obviously not the quality of life. (Stiglitz et al. 2009, pp. 8–9)

15.1 Introduction

The attention being given to well-being as a possible complement to the conventional measures of economic performance constitutes one of the notable turning points in our measurement of progress. However, the extent to which the ideas promulgated in the Stiglitz Report get translated into national policy remains to be seen (Stiglitz et al. 2009). Much will depend on our agreement over the measurement of both growth and well-being. At this stage, we are still seeking a consensus on the measurement of well-being and have only just begun assessing its value in a number of policy arenas. This is in spite of the listing of over 10,000 published papers on the World Database of Happiness since it was set up a decade ago.

Despite the impressive growth in what is routinely referred to as the “happiness” literature, relatively little attention has been paid to well-being in a regional context. A review of papers published in 17 regional science and related journals over the last 30 years uncovered less than two dozen papers. A broader search under regions in the World Database of Happiness revealed only 74 references. There has therefore only been a minor overlap between the concerns of regional scientists and those addressing happiness and well-being.

The limited material on well-being within regional science journals falls into four categories. The first involves the mapping and interpretation of distributions of well-being within and between regions. A second distinct subset of this literature identifies the way in which well-being varies by population density (urban versus rural). Another extends the regional convergence literature to embrace well-being (often comparing results to convergence in income). A fourth attempts to identify the way contexts, that is, the features that make places different, are associated with higher or lower well-being. However, except for environmental concerns, few studies have attempted to measure well-being returns to public investment in places.

The opportunities subjective well-being measures offer to regional scientists are unusual. The discipline’s twin concerns with the spatial dimension and public policy means that much of our concern for well-being derives from our interest in context, in the role that place plays in people’s livelihood. Over much of the history of regional science, we have sought to measure the impacts of increasing accessibility of locations, for example, via their effect on price by means of land rent.

By contrast, subjective well-being measures offer an opportunity to consider the effect of location (accessibility) and other changes to place more directly on people's well-being especially where impacts through the market are either not present or difficult to disentangle.

When it comes to individual well-being, the central question faced by the regional scientist is that posed by equilibrium theory – the idea that in the long run, people and places will mutually adjust so that well-being no longer varies geographically: see, for example, the standard Roback model (1982). Over time, people will move or adapt to the point where there is no further advantage to be gained from moving elsewhere. For a variety of reasons, including lack of information or the cost of mobility, the distribution of population may remain in a state of disequilibrium, with some people not making the adjustment from an existing suboptimal location to a better alternative. The relief map of well-being will therefore typically exhibit peaks and troughs.

People adjust their location as they do other aspects of their lives – in order to minimize tensions between aspirations and their realization (Clark et al. 1984). They do so by adapting as stayers or adjusting as movers. Any relief map of well-being will therefore reflect both frictional and structural pockets of low well-being. The challenge to regional scientists lies in separating out individual from context effects on well-being in ways that also take into account the dynamics of the adjustment process itself – at the level of both the individual and the society.

This chapter consists of six sections beginning with an introduction to the concept of well-being as a subjective measure. The literature's central question is raised in Sect. 15.3 – the relationship between regional economic growth and well-being. Section 15.4 reviews several examples of the way well-being has been analyzed from a regional perspective. Section 15.5 considers one of the main ways the subjective measures differs from objective measures of well-being, namely, the sensitivity of well-being to relative social position or rank and the implications this has for the way in which people cluster in space. The chapter ends with a summary of the conceptual and empirical challenges which subjective measures of well-being pose for regional science.

15.2 Well-Being

The literature on the “economics of happiness” has succeeded in establishing the broad demographic and other personal attributes that influence subjective well-being to the point where we can measure with some reliability not only the effects of material standards of living but also the marginal impact of “external shocks” on people's perception of their own well-being (Frey and Stutzer 2002; Kahneman and Krueger 2006). Changes in external conditions of general (Oswald 2003) or of particular events such as unemployment (Winkelmann 2009) are associated with changes in our level of well-being.

Drawing heavily on relevant literature in psychology, economists have now gained sufficient confidence in the properties of responses to questions on subjective well-being to use them as dependent variables in multivariate modeling contexts.

If modeled appropriately, responses to these questions can identify consistent correlations with respondent attributes like gender, age, and income. In doing so, the vast majority of studies within the economics of happiness literature employ a single metric of happiness – responses to the life satisfaction question: “All things considered, how do you feel about your life (right now)?” Possible answers are selected from an ordinal scale ranging from Very Satisfied to Very Dissatisfied.

One of the reasons for the rapid spread of research on well-being within economics in particular has been the fact that the outcome or dependent variable is relatively easy to collect and key parameters are straightforward to estimate plus the fact that there are now a wide range of sample surveys of individuals available in most countries. Although ordinal probit or logit models are technically the most appropriate approach to analyzing ordinal scales of this type, very similar results with more straightforward interpretations are frequently obtained using more familiar OLS regression estimation (Ferrer-i-Carbonell and Frijters 2004).

Readily available microdata from large surveys offer new opportunities for regional scientists to examine the way economic growth affects people’s well-being not just at the national level but also in specific regional settings. Surveys which leverage off natural experiments, for example, have allowed the influence of external circumstances on internal expressions of happiness to be estimated (Easterlin and Zimmermann 2008). The lessons from research at the level of the country as a whole suggest considerable potential for analysis at the regional level.

15.3 Economic Growth and Well-Being

One of the most influential stimulants to the growth of “happiness economics” has been the much referred to evidence that continued economic growth among affluent countries has failed to raise average levels of happiness. In referring to his seminal paper (1974), Richard Easterlin drew our attention to the fact that “At a point in time persons with higher income are, on average, happier than those with less; over time, however, as incomes increase generally in the course of modern economic growth, there is no improvement in the average level of happiness” (Easterlin 2011, p. 6). There is therefore a “distinction between the shorter term association between happiness and income, which is positive, and the longer term relation, which is nil” (Easterlin 2011, p. 7) – other things equal.

The key to the puzzle lies in the interplay between a subjective variable, material aspirations, and an objective one, household income. At a point in time the dispersion among individuals in material aspirations is less than in household income, and those with higher income are consequently better able to realise their aspirations, and are happier. Over time, however, material aspirations rise, on average, in about the same proportion as average income and undercut the increase in happiness that higher income would otherwise bring. (Easterlin 2011, p. 8)

In other words, while people adapt hedonically to an increase in income, their material aspirations tend to rise commensurately, but material aspirations are much

less flexible downward because once people have attained a given level of income, they cling to this reference point.

Developed originally for countries, there is no reason, except for data availability, why the same argument could not be applied to regions, that is, to particular subnational territories such as cities. Indeed, local expressions of the Easterlin paradox have already been suggested (Morrison 2011). The relationship between well-being and development, which has been debated at a global level, has a counterpart in the internal interaction between growth and well-being within countries. The transfer is not seamless, however, and there are some particular complications in transferring questions about well-being from countries to regions.

15.4 Subjective Well-Being and the Region

In his now classic 1981 book, *The Sense of Wellbeing in America*, Angus Campbell explored the impact of various domains of people's lives on their well-being (Campbell 1981). The domain of relevance was "where you live," and Campbell observed how subjective indicators of quality of life in the different regions departed dramatically from what he had learned from comparison of their objective characteristics (Campbell 1981, p. 147). The central point of difference was their variability. Regional differences in perceived well-being were modest at best and could not be explained by the much greater variability in economic and related characteristics of regions (Campbell 1981 148). This is of course what one would expect on the basis of spatial equilibrium theory.

The greater similarity regions exhibit in average subjective well-being compared to objective measures of well-being has been complemented by evidence showing that subjective well-being measures converge more quickly over time. Since Campbell wrote, there have been several studies which have showed the way in which broad quality of life indicators converge more quickly across regions than does income (Merchante and Ortega 2006), although convergence does not have to occur across all the indicators (Royuela and Artis 2006).

That regional differences in subjective measures of well-being show less variation than the objective conditions which partly give rise to them and converge more rapidly over time is consistent with the notion of adjustment undertaken through migration and adaption that underlies the concept of spatial equilibrium in well-being. However, the presence of persistent geographic pockets of low well-being raises important questions about the way in which differential economic growth distributes people spatially according to their ability to realize expectations either globally, as in overall well-being, or locally, as in domain-specific well-being.

Astra Bonini, for example, wanted to understand which of the national level indicators that are frequently used as proxy measures of development best reflect average life satisfaction and whether it was possible for universal indicators to capture such variations (Bonini 2008, pp. 224–225). "Understanding whether and how individual life satisfaction varies across countries is important because if the

goal of development is to increase wellbeing, we must identify the causes of wellbeing in different national and regional contexts.” (Bonini 2008, p. 223). She goes on to use hierarchical linear modeling of World Values Survey data to test not only whether individual well-being varies by country but also whether the same characteristics of individuals explain satisfaction across different countries. Different factors predict subjective well-being in different nations (Inglehart and Rabier 1986), as was discovered early on in the history of well-being research (Cantril 1965; Andrews and Withey 1976). The key research question is the source of this heterogeneity and the way country context influences the way individuals judge their own well-being.

Not only do different factors contribute to well-being in different countries but the same factors can also exert different levels of influence in different regional contexts. Bonini finds, for example, significant variations in the slopes of individual predictors of life satisfaction across both countries and regions (Bonini 2008). Context, she observes, acts as a filter: “aging one year in Latin America decreases life satisfaction more than aging one year in Europe” (Bonini 2008, p. 226); “individual income do(es) not increase life satisfaction as much in nations with relative high GDP per capita,” “increases in GDP per capita significantly decreases the positive effect that individual income has on life satisfaction,” and “unemployment in high income [countries] causes a greater reduction in life satisfaction than in [lower income] countries” (Bonini 2008, p. 231).

Just as Bonini, a sociologist, showed how country context filters the effect that individuals’ attributes have on well-being, so have economists Andrew Clark and colleagues. After comparing samples from European countries, they strongly reject the hypothesis that individuals transform income into well-being in the same way in different countries (Clark et al. 2005).

To the economist focused on individuals, place is often treated simply as a source of unobserved individual heterogeneity (in contrast to human geographers who have constructed a whole discipline around the presence of such heterogeneity). Two sources of heterogeneity are of specific interest in the well-being-income relationship. The first is intercept heterogeneity, which is the difference in the level of well-being at given incomes across regions. The second is slope heterogeneity, which is the difference in the degree to which utilities rise with income from one region to another.

From their European survey, Clark et al. find that different (endogenously identified) “classes” of individuals show very sharp differences in the effect of income on their declared satisfaction (Clark et al. 2005, p. 125). “The marginal effect of income on subjective wellbeing depends,” they explain, “on unobserved heterogeneities relating either to the underlying utility function or to the way people label their utility” (Clark et al. 2005, p. 127). One group, they find, is both highly satisfied and has large marginal effects of income on well-being (the older, more affluent but less active in the labor market), while another is the least satisfied and has the lowest marginal effects of income on wellbeing (the less well educated, the unemployed and single) (Clark et al. 2005, p. 127). Northern European countries are overrepresented in the former and southern European countries in the latter.

Using a similar methodology, John Helliwell notes how “life satisfaction rises less with relative income in the OECD than the non-OECD countries,” and how female advantage is larger in OECD countries due to higher status of women (Helliwell 2008, p. 6). Applying multilevel modeling to subjective well-being measures across the European Union, Pittau et al. also point to the role of context in interpreting income effects, noting in particular how personal income matters more in poor regions than in rich regions (Pittau et al. 2010, p. 341). They find that, even after controlling for individual characteristics (and different effects of income and employment status across regions), the unexplained regional-level variability of the estimated life satisfaction is still high, “indicating that geography still matters considerably” (Pittau et al. 2010, p. 358).

In another study, devoted to the British experience, Ballas and Tranmer asked whether levels of happiness among individuals in Britain reflect different characteristics of residents in different districts, regions, and areas (compositional effects) or whether there are environmental or other factors such as social capital, cohesion, and socioeconomic inequality (contextual effects) that cause their inhabitants to be happy or unhappy (Ballas and Tranmer 2012, p. 80). Such issues also attract the attention of health geographers and epidemiologists with respect to mental health and morbidity.

Ballas and Tranmer find only weak contextual effects, a result which can surprise geographers conditioned to focus on differences across space. However, this is precisely the result economic theory would suggest. People move to the places that suit them best and move again if either their own circumstances or the characteristics of the place they live in changes away from an acceptable degree of harmony. Therefore, while at any point in time there will be a set of individuals dissatisfied with their place, this disconnect is likely to be primarily frictional. After controlling for the characteristics of individuals themselves, one would not expect persistent differences in the contribution place makes to subjective well-being across regions, especially in countries like Britain or the United States that are used to high levels of interregional migration.

However, such a spatial equilibrium argument only applies to those who are free and able to move and have reasonable knowledge of alternative locations. Excluded from this assumption are at least three population groups whose ability to escape less than optimal locations is dependent on others: children, the unemployed, the poor and the elderly. One of the challenges in estimating regional context effects therefore lies in measuring the greater sensitivity of the less mobile to context effects.

While recent research has drawn attention to the way in which geographical context affects the way attributes of individuals are reflected in assessments of their own well-being, only recently has it been explicitly recognized that contexts themselves not only have scale but neighboring properties. Using NUTS level 1 regions for Europe, Okulicz-Kozaryn (2011), for example, has shown that happy regions tend to be surrounded by other happy regions, and unhappy regions tend to be surrounded by unhappy regions, in what he has identified in Europe as happiness clusters.

Another example of systematic effects of context on well-being has emerged in the study of population density, or urbanization, on well-being. Here, there are three distinct perspectives. The first argues that economic growth closes (or in some cases reverses) the gap between well-being in urban and rural areas. The second explores the cross-sectional negative relationship between well-being and settlement density, and the third, more disparate, literature attempts to identify those attributes of cities that actually generate variations in well-being.

In the first case, Easterlin et al. (2011) view spatial differences in well-being as the geographical expression of the increasing marginal utility of income: as some places become richer, their average level of well-being rises.

At low levels of economic development there are substantial gaps favouring urban over rural areas in income, education and occupational structure, and consequently a large excess of urban over rural life satisfaction, despite important urban problems of pollution, congestion, and the like. At more advanced development levels, these economic differentials tend to disappear, and rural areas approach or exceed urban in life satisfaction (Easterlin et al. 2011, p. 2189)

There are now many examples in which the largest cities in an urban hierarchy now report relatively low levels of life satisfaction: Dublin, “inner London,” the largest US cities, Sydney, and Auckland City – even after controlling for both individual and, in some cases, neighborhood fixed effects. However, the relationship between population density and well-being may be more general than the focus on primate cities suggests. Berry and Okulicz-Kozaryn (2011), for example, identify a gradient of subjective well-being (happiness) that rises from its lowest levels in large central cities to its highest levels on the small town/rural periphery (at least in the USA). In response to those documenting the manifold economic returns to increasing city size, they show that while there are many benefits of big-city living, high levels of happiness are not among them (Berry and Okulicz-Kozaryn (2011), p. 872).

The geography of well-being may not mirror the geography of growth. As Campbell (1981) observed much earlier on, “Despite . . . widely proclaimed failings of urban life, Americans have stubbornly persisted in migrating from the rural areas into the nation’s towns and cities.” People went to the cities because that is where the jobs were. Had they had their choice, they might have preferred to remain in their rural setting, but the exigencies of making a living compelled them to move to the city” (Campbell (1981), p. 149). Metropolitan people are therefore, “most inclined to believe that they have not had their full share of happiness in life” (Campbell (1981), p. 150).

A third approach to well-being in the regional literature has tried to identify those attributes of settlements that actually contribute to well-being. Here Marans and Stimson (2011) have built on the seminal work of Campbell in arguing that the quality of a place is a subjective phenomenon and therefore can vary across individuals, their demographic and socioeconomic characteristics, as well as their past experiences and aspirations.

In summary, the focus of well-being research in economics has been either on the way in which average levels of individual well-being rise with economic

growth or the way well-being declines without it. The finding that the local engines of growth, the large agglomerations, are those least likely to return high levels of well-being within their localities is a salient reminder that objective and subjective measures of well-being can diverge. The reasons for this divergence and the contexts in which they take place are both research frontiers.

One such reason has to do with the way the subjective is sensitive to relativities. Lying behind the Easterlin paradox, for example, is the argument that people compare themselves to others, and the closer they are both personally and geographically, the more influential the comparison is on subjective reports of well-being. Interdependency in preferences has an important distance decay property, and this is why comparisons between countries show an association between income and happiness which is so much weaker than those comparisons that take place within countries (Easterlin 2011, p. 35). The spatial properties of interdependencies in preferences may be particularly significant in understanding subjective well-being. As the positional goods literature suggests, consumption behaviors reflect this feature in ways that have yet to be fully exploited by the regional scientist (Cheshire et al. 2003; Durlauf 1996; Frank 2005).

15.5 Well-Being and Relativities

One of the challenges that regional science faces is how to handle the spatial properties of the comparisons that appear so influential in the measurement of subjective well-being. As is well established in the workplace and in residential location decisions, relative incomes and associated social status have a direct bearing on people's assessment of their own well-being (Marmot 2005).

One of the advantages of taking a subjective approach to well-being is that it allows us to view the residential location decisions as a way of resolving tensions arising out of relative status. This feature of residential location remains largely untouched in the conventional bid rent curve interpretation of location by income group in part because the interdependences of preferences remained unrecognized in such models (Duesenberry 1949). Those that have been developed, using agent-based modeling, for example, show how even mild preferences for association can lead, as Schelling argued, to spatial patterns of segregation (Clark and Fossett 2008; cf Schelling 1978). Cellular automata simulations have also shown how the attraction of cheaper goods used by neighbors can lead to a spatial clustering of consumption patterns (Bell 2002).

Interdependencies in consumption highlight the fact that one person's utility is not independent of others. Since people are sensitive to the social status of others, their own status may well increase by adopting consumption patterns of neighbors or by moving to areas where they occupy higher relative rank and they can demonstrate their higher *relative* income (expenditure). However, individuals look up but not down when making comparisons (Duesenberry 1949, pp. 234–235), and empirically, there is little evidence that individuals prefer low-income reference groups as a way of "self-enhancement"; instead they are more likely to move to areas where the gap between them and the next up is minimized.

According to this second argument, people perform better and are more successful if they set themselves high goals or compare with high reference standards (Clark et al. 2008, p. 113). The fact that social externalities impinge asymmetrically on individuals' well-being and behavior has influenced the way we think about the spatial adjustments people make including where they migrate (Stark and Wang 2005). Reference group choice can therefore be used to explain migration decisions of heterogeneous individuals: "What happens . . . when people who care about their relative position in a group have the option to react by staying in the group or exiting from it?" (Stark and Wang 2005, p. 223).

Stark and Wang go on to describe an endogenous process of voluntary segmentation, although based explicitly on relative social rank rather than preference per se (cf Schelling 1978). When individuals who initially belong to one group act (costlessly) upon their distastes for relative deprivation and self-select into any one of two groups, they end up splitting into two groups in a manner that is sensitive to the way in which relative deprivation is sensed and measured (Stark and Wang 2005, p. 233). By this process, aggregate relative deprivation is lowered, and social welfare is enhanced (largely due to a reduction of social tensions).

While looking up may dominate, people do also look down, and this provides a useful link back to the relationship between relativities of consumption and the spatial clustering of consumption levels. As Stark and Wang note, "while the utility of an individual rises in his own consumption, it declines in the consumption of any of his neighbours if that consumption falls below some minimal level" (Stark and Wang 2005, p. 235). In other words, individuals are adversely affected by the material well-being of others in their reference group when this well-being is sufficiently lower than theirs.

Placing subjective measures of well-being in geographical context requires regional scientists to consider relativities and to do so in a much more spatially sensitive way than Easterlin was required to do at the country level. What remains conceptually problematic is the adjustment path people take when they face discomfiting levels of social relativity. People may be prepared to undergo short-term loss of well-being as they move closer to those some social distance above them in the hope that this proximity will eventually raise them up. A willingness to experience such temporary initial discomfort may help explain why well-being tends to be lower in very large centers where income inequalities are wider but where chances of upward mobility are also greater, or believed to be so. Another possibility is that if countries as a whole do not become happier as they become richer, individuals may still continue to adjust where they live within the country (or city) in order to extract greater status out of the increases in income that occur to them personally. We might ask therefore whether the search for greater well-being as subjectively measured within affluent economies is being increasingly sought *within* countries through spatial segregation?

This same issue of relativities suggests that we ask whether, as they become richer, people are more likely to spatially cluster in order to exploit the externalities from positional goods and whether, as a result, the poor need to cluster more in order to compensate (Durlauf 1996). This involves the deeper welfare question

involving the longer run polarizing effects of residential segregation – the possibility that residence in well-endowed areas on one hand and deprived areas on the other actually inhibits social mobility of the latter, thus widening the gap between rich and poor (Clark and Morrison 2012).

In summary, considering well-being as a subjectively generated measure opens up avenues for assessing the influence of economic growth on localities that have scarcely been addressed by regional scientists. The way in which sensitivities to social rank can lead to residential location decisions, for example, runs the possibility of increasing levels of spatial inequality in ways that might not be captured through the usual market signals.

15.6 Conclusions

If the development of policy instruments lags behind the flurry of attention paid to well-being in general, then the gap is considerably wider when it comes to understanding the link between *regional* growth and changes in local well-being. This is not because the spatial distribution of well-being has not received attention, for there are multiple papers on this topic, but most are motivated by the literature on social indicators rather than the economics of happiness. The latter is not focused on the spatial distribution of well-being per se but in how sensitive changes in well-being are to changes in income – on whether economic growth increases happiness at both the macro and individual levels.

What the regional literature has yet to grapple with is how the relationship between income and well-being works out locally both within and across regions. We do not know enough about how people adapt to suboptimal locations nor how sensitive people's mobility is to their perceptions of their own well-being. However, approaching the location issue from the perspective of subjective well-being throws into sharp relief the role of social relativities or rank. Therefore, one of the major differences between the study of well-being subjectively and at the regional level is the greater importance of the local reference group. At the national level, reference groups are treated as national norms, but once people are anchored locally by their spatial subscripts, the local frames of reference become important. One of the spatial issues the attention to subjective well-being raises is how responses to local reference groups are related to socio-spatial segregation. So far little attention has been paid to the way in which socio-spatial segregation is used as a way of regulating happiness levels locally.

Central to any such advance is how we measure well-being itself. After several decades of well-being research, there is a growing realization, particularly now that policy applications are being actively considered, that if we are to judge the impact of external events on place, then our measurement of well-being needs to become more sophisticated. It is now well established that there is a difference between happiness as a measure of everyday emotions, and happiness as satisfaction with life; the former addresses everyday moods, while the latter is more reflective of a longer time span. A third dimension has to do with meaning, the purposive aspect

of life, often referred to as eudaemonic happiness (from the Greek word *eudaimonia*). It may even be necessary to capture the entire range of dimensions by relying on five separate questions (Graham 2011, p. 37): “the ladder-of-life question (Cantril 1965), the life satisfaction question, two questions to measure experienced utility (for example, both positive affect, as indicated by smiling, and negative affect, as indicated by worry), and a question that captures life purpose (Dolan et al. 2008).”

How we measure people’s expression of their own well-being has received considerable attention in positive psychology, and this chapter has asked how the study of well-being from this subjective perspective might contribute to regional science. A second and somewhat more challenging question is how regional science can now contribute to the study of well-being.

References

- Andrews FM, Withey SB (1976) Social indicators of wellbeing: American’s perceptions of life quality. Plenum Press, New York
- Ballas D, Tranmer M (2012) Happy people or happy places? A multilevel modeling approach to the analysis of happiness and wellbeing. *Int Reg Sci Rev* 35(1):70–102
- Bell AM (2002) Locally interdependent preferences in a general equilibrium environment. *J Econ Behav & Organ* 47(3):309–333
- Berry BJL, Okulicz-Kozaryn A (2011) An urban–rural happiness gradient. *Urban Geogr* 32(6):871–883
- Bonini AN (2008) Cross-national variation in individual life satisfaction: effects of national wealth, human development, and environmental conditions. *Soc Indic Res* 87(2):223–236
- Campbell A (1981) The sense of wellbeing in America. McGraw Hill, New York
- Cantril H (1965) The pattern of human concerns. Rutgers University Press, New Brunswick
- Cheshire PC, Monastiriotis V, Sheppard S (2003) Income inequality and residential segregation: labour market sorting and the demand for positional goods. In: Martin R, Morrison PS (eds) *Geographies of labour market inequality*. Routledge, London, pp 83–109
- Clark WAV, Fossett M (2008) Understanding the social context of the Schelling segregation model. *Proc Natl Acad Sci* 105(11):4109–4114
- Clark WAV, Morrison PS (2012) Socio-spatial mobility and residential sorting: evidence from a large-scale survey. *Urban Stud* 49(15):3253–3270
- Clark WAV, Deurloo MC, Dieleman FM (1984) Housing consumption and residential mobility. *Ann Assoc Am Geogr* 74(1):29–43
- Clark A, Etilé F, Postel-Vinay F, Senik C, Van der Straeten K (2005) Heterogeneity in reported wellbeing: evidence from twelve European countries. *The Econ J* 115:C118–C132
- Clark AE, Frijters P, Shields MA (2008) Relative income happiness, and utility: an explanation for the Easterlin paradox and other puzzles. *J Econ Lit* 46(1):95–144
- Dolan P, Peasgood T, White D (2008) Do we really know what makes us happy? A review of the literature on the factors associated with subjective wellbeing. *J Econ Psychol* 29(1):94–122
- Duesenberry JS (1949) Income, saving and the theory of consumer behaviour. Harvard University Press, Cambridge, Massachusetts
- Durlauf SN (1996) A theory of persistent income inequality. *J Econ Growth* 1:75–79
- Easterlin R (1974) Does economic growth improve the human lot? Some empirical evidence. In: David PA, Melvin WB (eds) *Nations and households in economic growth*. Stanford University Press, Palo Alto, pp 89–125

- Easterlin RA (2011) Happiness, growth and the life cycle. Oxford University Press for IZA, Oxford
- Easterlin RA, Zimmermann AC (2008) Life satisfaction and economic conditions in East and West Germany pre- and post-unification. In Working paper SOEP (DIW, Berlin, Germany)
- Easterlin RA, Angelescu L, Zweig JS (2011) The impact of modern economic growth on urban-rural differences in subjective wellbeing. *World Dev* 39(12):2187–2198
- Ferrer-i-Carbonell A, Frijters P (2004) How important is methodology for the estimates of the determinants of happiness? *The Econ J* 114(497):641–659
- Frank RH (2005) Positional externalities cause large and preventable welfare losses. *Am Econ Rev* 95(2):137–141
- Frey BS, Stutzer A (2002) Happiness and economics: how the economy and institutions affect human well being. Princeton University Press, Princeton
- Graham C (2011) The pursuit of happiness: an economy of wellbeing. Brookings Institution Press, Washington
- Helliwell JF (2008) Life satisfaction and the quality of development. Working paper 14507, National Bureau of Economic Research, Cambridge, MA
- Inglehart R, Rabier JR (1986) Aspirations adapt to situations – but why are the Belgians so much happier than the French? In: Andrews FM (ed) Research on the quality of life. Institute for Social Research, University of Michigan, Ann Arbor, pp 1–56
- Kahneman D, Krueger AB (2006) Developments in the measurement of subjective wellbeing. *J Econ Perspect* 20(1):3–24
- Marans RW, Stimson RJ (2011) Investigating quality of urban life: theory, methods, and empirical research. Springer, London/New York
- Marmot M (2005) The status syndrome: how social standing affects our health and longevity. Owl Books, New York
- Merchante AJ, Ortega B (2006) Quality of life and economic convergence across Spanish regions, 1980–2001. *Region Stud* 40(5):471–483
- Morrison PS (2011) Local expressions of subjective wellbeing: the New Zealand experience. *Region Stud* 45(8):1039–1058
- Okulicz-Kozaryn A (2011) Geography of European life satisfaction. *Soc Indic Res* 101(3):435–445
- Oswald AJ (2003) How much do external factors affect well being? *Psychologist* March 16(3):140–141
- Pittau MG, Zelli R, Gelman A (2010) Economic disparities and life satisfaction in European regions. *Soc Indic Res* 96(2):339–361
- Roback J (1982) Wages, rents and the quality of life. *J Political Econ* 90(6):1257–1278
- Royuela V, Artis M (2006) Convergency analysis in terms of quality of life in the urban systems of the Barcelona province, 1991–2000. *Region Stud* 40(5):485–492
- Schelling TC (1978) Micromotives and macrobehaviour. Norton, New York
- Stark O, Wang YQ (2005) Towards a theory of self-segregation as a response to relative deprivation: steady-state outcomes of social welfare. In: Bruni L, Porta PL (eds) Economics and happiness: framing the analysis. Oxford University Press, Oxford/New York, pp 223–242
- Stiglitz JE, Sen A, Fitoussi JP (2009) Report by the commission on the measurement of economic performance and social progress. (The Commission on the Measurement of Economic Performance and Social Progress (CMEPSP))
- Winkelmann R (2009) Unemployment, social capital, and subjective wellbeing. *J Happiness Stud* 10(4):421–430

Julie Le Gallo and Bernard Fingleton

Contents

16.1	Introduction	292
16.2	Growth Regressions: From Theory to Empirics	292
16.3	Estimating the Rate of Convergence	299
16.3.1	Unconditional and Conditional β -Convergence	300
16.3.2	Space and Growth	301
16.3.3	Econometric Issues	302
16.3.4	Panel Estimation	304
16.3.5	Multiple Regimes and Convergence Clubs	307
16.4	Sigma-Convergence and Distribution Approach to Convergence	308
16.4.1	σ -Convergence	308
16.4.2	Studying the Evolution of the Cross-Sectional Distributions	309
16.4.3	Distribution Dynamics and Space	312
16.5	Conclusions	313
	References	314

Abstract

This chapter provides a selective survey of the main developments related to the study of regional convergence. We discuss the methodological issues at stake and show how a number of techniques applied in cross-country studies have been adapted to the study of regional convergence. In doing this, we focus on the two main strands of growth econometrics: the regression approach where

J. Le Gallo (✉)

CRESE, Université de Franche-Comté, Besançon, France
e-mail: jlegallo@univ-fcomte.fr

B. Fingleton

Department of Economics, University of Strathclyde, Glasgow, Scotland, UK
e-mail: bf100@cam.ac.uk

predictions from formal neoclassical and other growth theories have been tested using cross-sectional and panel data and the distribution approach, which typically examines the entire distribution of output per capita across regions. In each case, we show how the analysis of regions rather than countries emphasizes the need to take proper account of spatial interaction effects.

16.1 Introduction

Given the persistent disparities in aggregate growth rates between countries and even within countries, the question whether incomes are converging across regions has received a lot of attention in the last two decades. From a theoretical point of view, regional growth modeling has been largely motivated by work done at the cross-country level, notably by Barro (1991), Barro and Sala-i-Martin (1995), and Mankiw et al. (1992), who developed empirical models based on the Solow-Swan economic growth model. These neoclassical models have as a major prediction the convergence of countries or regions to an equilibrium at which growth settles down to a constant rate, referred to as the steady state. Set against this are numerous variants on the basic theory and more radical departures from neoclassical principles, which allow non-convergent outcomes.

This chapter provides an overview of the main developments related to the study of regional convergence. We discuss the methodological issues at stake and show how a number of techniques applied in cross-country studies have been adapted to the study of regional convergence. In doing this, we focus on the two main strands of growth econometrics: the regression approach where predictions from formal neoclassical and other growth theories have been tested using cross-sectional and panel data and the distribution approach, which examines the entire distribution of regions. In each case, we show how the analysis of regions rather than countries emphasizes the need to take proper account of spatial interaction effects.

The chapter is organized as follows. In Sect. 16.2, we present a simple theoretical framework for two regions describing the neoclassical growth model. Section 16.3 provides a survey on the regression approach based on the concept of β -convergence and its spatial extensions. Section 16.4 examines the distribution dynamics approach together with exploratory spatial data analysis techniques. Section 16.5 concludes.

16.2 Growth Regressions: From Theory to Empirics

Consider two regions, each of which is governed by the same production technology, although there are differences between the regions, which lead them to separate parallel growth paths. The production technology can be described as

$$Y_{jt} = K_{jt}^\alpha (A_t H_{jt})^\beta \quad (16.1)$$

in which Y_{jt} is the level of output (GDP) in region j at time t , K_{jt} denotes the level of capital in region j at time t , A_t is labor augmenting technology (total factor productivity), and H_{jt} is the level of skilled labor. Dividing variables on both sides by $A_t H_{jt}$, we have output and capital per unit of effective labor:

$$\tilde{y}_{jt} = \tilde{k}_{jt}^\alpha \quad (16.2)$$

where $\tilde{y}_{jt} = Y_{jt}/A_t H_{jt}$ and $\tilde{k}_{jt} = K_{jt}/A_t H_{jt}$. In writing this, we assume that $\alpha + \beta = 1$, that is, constant returns to scale, with capital's share of income equal to α and augmented labor's equal to $1 - \alpha$, with diminishing returns to capital and augmented labor.

Consider now the dynamics entailed by this model. First, we assume that technology A grows at the constant rate g and raw labor L grows at the rate n_1 in region 1 and n_2 in region 2. For the moment, this is the only difference assumed between the regions. Second, assume that skilled labor H is determined by the years of schooling (c) and the rate of return per year of schooling (ϕ) that raw labor experiences. The product $c\phi$ determines the rate at which raw labor turns into skilled labor. Finally, the level of capital K is determined by the investment rate I and the depreciation rate d of existing capital, with investment equal to a share s of output Y . We capture the dynamics with the following system:

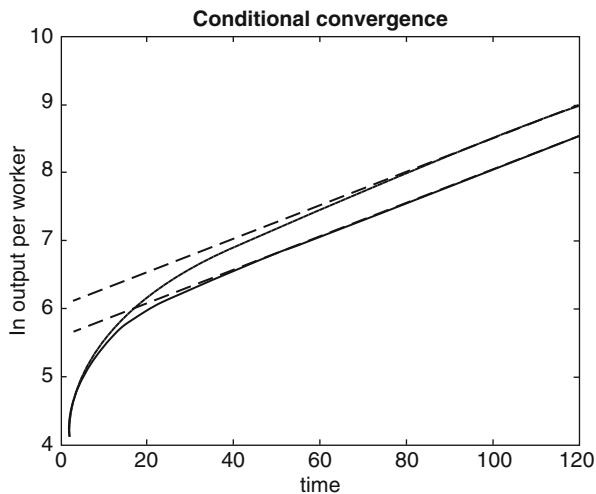
$$\begin{aligned} A_t &= A_{t-1} (1 + g) \\ L_{jt} &= L_{jt-1} (1 + n_j) \\ H_{jt} &= L_{jt} (1 + \phi c) \\ K_{jt} &= I_{jt-1} (1 - d) K_{jt-1} \\ I_{jt} &= s Y_{jt} \end{aligned} \quad (16.3)$$

[Figure 16.1](#) shows the evolution of the system based on some assumptions (for visual effect rather than realism) about initial values and parameters α , g , n_1 , n_2 , $c\phi$, d , and s . We assume that $g = 0.025$, $s = 0.5$, $A = 110$, $K = 88.875$, $L = 20$, $Y = 90$, $c = 9$, $\phi = 0.1$, $\alpha = 0.333$, $d = 0.025$, $n_1 = 0.01$, and $n_2 = 0.1$. While both regions start from the same position, they move onto different steady-state paths of growth in output per worker as a result of their differing labor force growth rates.

Convergence to equilibrium is determined by the fundamental assumption of the neoclassical growth model that there are diminishing returns. To show this, consider the derivative of output per unit of effective labor with respect to capital per unit of effective labor:

$$\begin{aligned} \frac{\partial \tilde{y}_{jt}}{\partial \tilde{k}_{jt}} &= \alpha \tilde{k}_{jt}^{\alpha-1} > 0, & \lim_{\tilde{k} \rightarrow 0} \left[\frac{\partial \tilde{y}_{jt}}{\partial \tilde{k}_{jt}} \right] &= \infty, \quad \lim_{\tilde{k} \rightarrow \infty} \left[\frac{\partial \tilde{y}_{jt}}{\partial \tilde{k}_{jt}} \right] &= 0 \\ \frac{\partial^2 \tilde{y}_{jt}}{\partial \tilde{k}_{jt}^2} &= (\alpha - 1) \alpha \tilde{k}_{jt}^{\alpha-2} < 0 \end{aligned} \quad (16.4)$$

Fig. 16.1 Conditional convergence for two regions



The first derivative is positive but goes to 0 as $\tilde{k}_{jt} \rightarrow \infty$, indicating that although the marginal product of capital is positive, capital deepening in the form of additional amounts of capital produces a diminishing rate of return (these are the Inada conditions).

The steady state to which the economy evolves is determined by the fact that although increasing income produces increasing investment, as shown by $I_{jt} = sY_{jt}$, there is a simultaneously occurring increase over time in aggregate depreciation, the most obvious component of which is due to capital depreciation, but which also depends on the growth in the effective number of workers. Moreover, while depreciation per effective worker is linear in capital per effective worker, investment is nonlinear, reflecting the diminishing marginal product of capital. This is shown in Fig. 16.2, which is the outcomes if we run our model and plot investment $sY_{jt}/(A_t H_{jt})$ (solid line) and depreciation per effective worker $(n_j + g + d)K_{jt}/(A_t H_{jt})$ (dotted line) against capital per effective worker $K_{jt}/(A_t H_{jt})$ using the data for region $j = 1$. Figure 16.2 shows that at low levels of capital per effective worker, investment is at a higher level than “depreciation.” However, with diminishing returns, the gap between the investment and depreciation schedule narrows progressively to the point where all savings are absorbed offsetting the effects of depreciation and effective labor force growth. Beyond this point, although additional income would generate additional savings and investment, the curvilinear savings schedule is now below the linear depreciation schedule, and the change in capital per effective worker becomes negative, and so the system moves back in the direction of falling income toward the equilibrium point. Thus, we have a stable equilibrium at which investment is just sufficient to balance the effects of depreciation and effective labor force growth and maintain the level of capital per effective worker.

Fig. 16.2 Investment vs capital per effective worker; First region

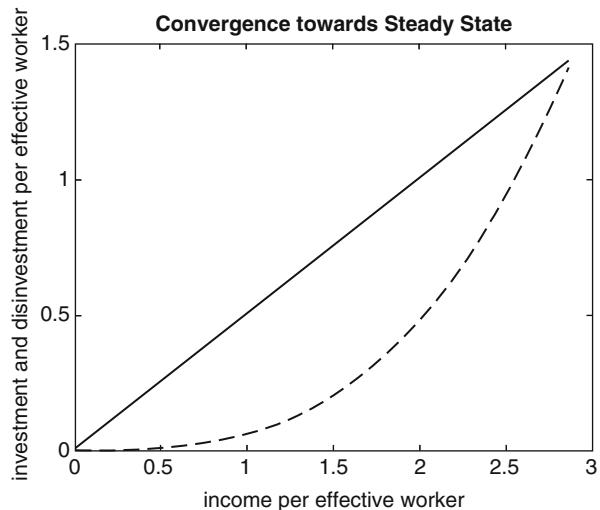
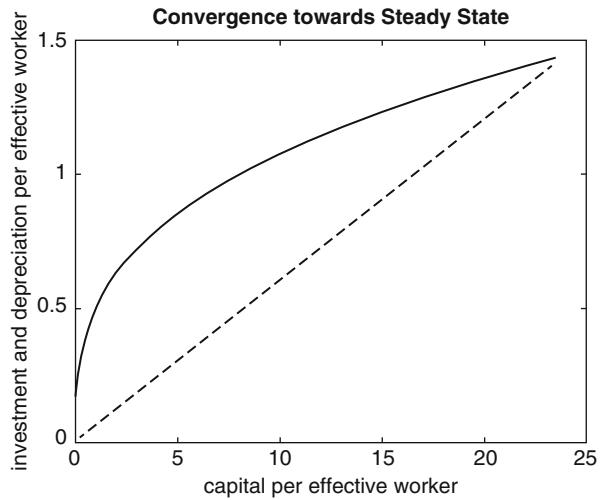


Fig. 16.3 Investment vs income per effective worker; First region

Figure 16.3 plots the same data but with income per effective worker $Y_{jt}/(A_t H_{jt})$ as the horizontal axis. Thus, using the data for region 1, this identifies the stable equilibrium point for income per effective worker as 2.86. Figure 16.4 is the equivalent data for the second region. Here, we see the effect of faster labor force growth, which produces a lower equilibrium point at about 1.81.

Figure 16.5 plots the two components of the right-hand side of the equation showing how capital per effective worker evolves, which is equal to

$$\dot{\tilde{k}}_t = s\tilde{y}_t - (n + d + g)\tilde{k}_t \quad (16.5)$$

Fig. 16.4 Investment vs income per effective worker; First region

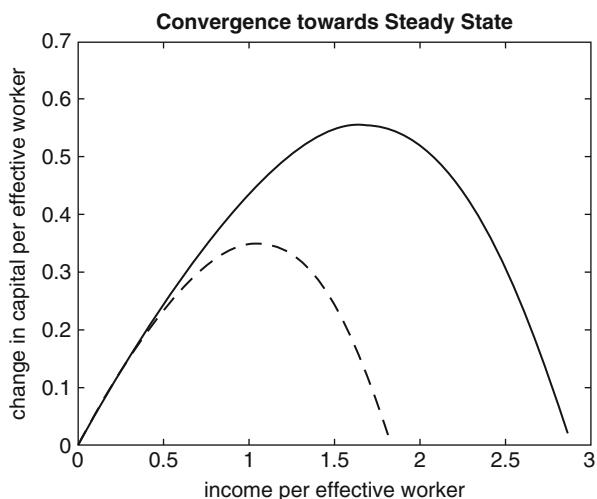
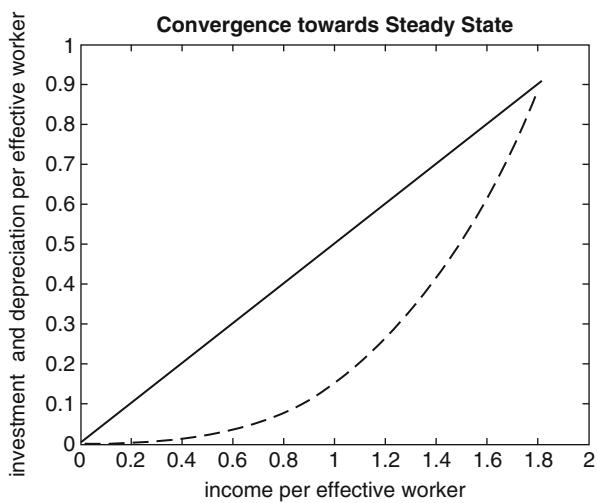


Fig. 16.5 Change in capital per effective worker

where $\dot{\tilde{k}}_t$ is the derivative of \tilde{k}_t with respect to time. From this, it is possible to obtain the equilibrium point equal to $\dot{\tilde{k}}_t = 0$, so that as we have shown graphically $s\tilde{y}_t = (n + d + g)\tilde{k}_t$. Figure 16.5 shows the evolution of $\dot{\tilde{k}}_t$ identifying our two equilibrium income per effective worker points at which $\dot{\tilde{k}}_t = 0$ for our two regions.

It follows that at equilibrium

$$s\tilde{y}_j^* = (n_j + d + g)\tilde{k}_j^* \quad (16.6)$$

Hence,

$$\begin{aligned}\tilde{k}_j^{*\alpha} &= (n_j + d + g) \tilde{k}_j^* \\ \tilde{k}_j^* &= \left[\frac{s}{n_j + d + g} \right]^{\frac{1}{1-\alpha}}\end{aligned}\quad (16.7)$$

and the equilibrium output per effective worker is

$$\tilde{y}_j^* = \tilde{k}_j^{*\alpha} = \left(\frac{s}{n_j + d + g} \right)^{\frac{\alpha}{1-\alpha}} \quad (16.8)$$

This means that equilibrium output is

$$Y_{jt}^* = \left(\frac{s}{n_j + d + g} \right)^{\frac{\alpha}{1-\alpha}} A_t H_{jt} \quad (16.9)$$

and equilibrium output per worker is

$$\frac{Y_{jt}^*}{L_{jt}} = \left(\frac{s}{n_j + d + g} \right)^{\frac{\alpha}{1-\alpha}} \frac{A_t H_{jt}}{L_{jt}} \quad (16.10)$$

Hence we have

$$\ln y_{jt}^* = \ln A_t + \frac{\alpha}{1-\alpha} \ln s - \frac{\alpha}{1-\alpha} \ln(n_j + d + g) + \ln \left(\frac{H_{jt}}{L_{jt}} \right) \quad (16.11)$$

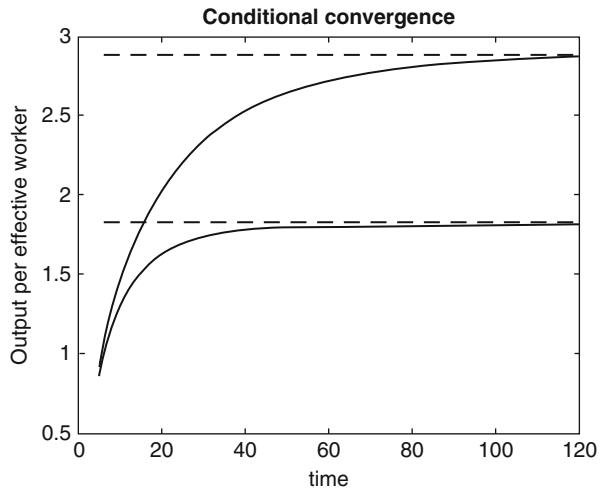
Equation (16.11) provides the equilibrium level of output per worker as traced by the broken lines of Fig. 16.1 for our two regions. It shows a steady growth, at rate g , but with different levels at any one point in time on account of the different labor force growth rates. In terms of output per unit of effective labor, we have seen from Fig. 16.5 and earlier that this converges to a constant 2.86 for region 1 and 1.81 for region 2. Following Eq. (16.10), the evolution toward this steady state is given by the constant

$$\ln \left(\frac{Y_{jt}^*}{A_t H_{jt}} \right) = \frac{\alpha}{1-\alpha} \ln s \frac{\alpha}{1-\alpha} \ln(n_j + d + g) \quad (16.12)$$

This is illustrated by Fig. 16.6.

We have given a highly stylized account of the determinants of regional growth, with regional differences existing purely as a consequence of differences in the rate of growth of labor. Thus, we have assumed that depreciation, returns to scale, the rate of technical progress, initial levels of technology, skilled labor, capital, and the savings rate are equal across our regions. Nevertheless, we see that this simple

Fig. 16.6 Evolution of output per effective worker



difference has consequences for the equilibria to which each region converges and the rate of convergence.

There is much interest in estimating convergence rates. As a result of linearizing the steady-state dynamics using a Taylor series expansion, we find that, approximately, the growth of output per effective worker is given by the gap between log level of output per effective worker and the log equilibrium level, thus

$$\frac{\partial \ln(\tilde{y}_{jt})}{\partial t} = -(1 - \alpha)(n_j + d + g)(\ln(\tilde{y}_{jt}) - \ln(\tilde{y}_{jt}^*)) \quad (16.13)$$

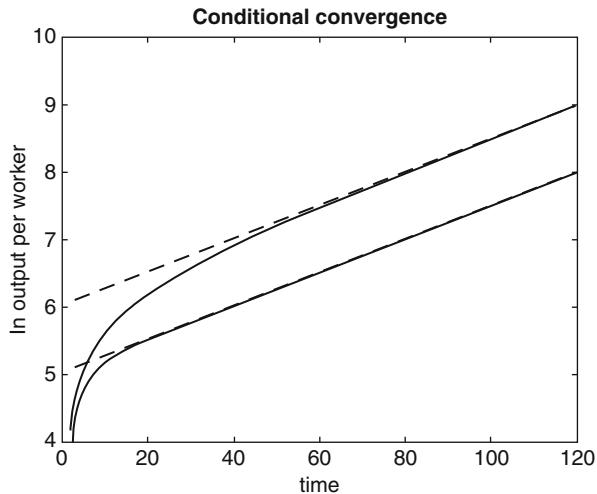
where the rate of convergence is $\beta_j = (1 - \alpha)(n_j + d + g)$. Note that for the parameters values in our example, $\beta_1 = 0.04$ and $\beta_2 = 0.1$, which compares with $\beta = 0.02$ (the so-called 2 percent rule) suggested by Barro and Sala-i-Martin (1995) which has in fact been observed in many growth studies. Integrating and writing in per worker terms, we obtain

$$\frac{1}{T} \ln\left(\frac{Y_{jt}}{L_{jt}}\right) = k - \frac{1}{T} e^{-\beta_j T} \ln\left(\frac{Y_{jt-T}}{L_{jt-T}}\right) \quad (16.14)$$

With large $\beta_j T$ the left-hand side is equal to k , which is proportional to the equilibrium level of output per worker.

One interesting prediction from the neoclassical growth model is the phenomenon of “catching up.” Consider two regions starting from different levels of output per worker. If we keep the equilibrium growth path for each the same for simplicity, then we find that there is faster growth in the initially poorer region. However, the prediction is more complex when both starting level and equilibrium path are different, as is more likely in the real world, as a result, for example, of a lower level of capital endowment and faster labor force growth rate. In our simulation,

Fig. 16.7 Convergence for two regions



the initially poorer country experiences a short-lived spurt of possibly faster growth at the outset, but over the longer term, we see growth moving sooner onto the equilibrium growth path entailing a lower equilibrium level of output per worker (obtained by setting $K = 1$ and $n_2 = 0.4$). [Figure 16.7](#) illustrates this outcome.

It should be noted that in the neoclassical growth model described above, there is no explicitly modeled spatial interaction between the regions, although the common growth rate of technology g may be interpreted as implying perfect diffusion of technological change across the regions. The model also assumes that the investment share of output is fixed over time. Once regional differences in investment behavior, innovation diffusion, and interregional migration are taken into account, the extent of catching up will strongly depend on the strength of these types of spatial interaction (see Nijkamp and Poot [1998](#)).

16.3 Estimating the Rate of Convergence

In this section, we review the main econometric issues associated with the estimation of the rate of convergence.

The debate on convergence has given rise to numerous empirical studies with often contradictory results due, partially, because various conceptions of convergence were tested and because various methodological approaches and procedures of tests have been used (cross section, panel data, temporal series, etc.). The first developments concern the idea of convergence-catching up, which is associated with the concept of β -convergence. This is based on the relationship between initial output and subsequent growth.

There are however two main approaches allowing the test of this hypothesis: absolute β -convergence and conditional β -convergence. Take again Eq. [\(16.11\)](#), which gives the equilibrium output per worker. This level depends upon several

parameters $\theta = (g, n, c, \phi, d, s, \alpha)$. If all elements of θ are similar for all regions, which then only differ by their initial effective per worker capital, then there is absolute β -convergence. If some elements of θ differ between regions, as was the case in our simulations, then there is conditional β -convergence.

16.3.1 Unconditional and Conditional β -Convergence

Consider first the simplifying case where all regions are structurally identical and have access to the same technology. They differ only by their initial conditions. In this case, they converge toward the same steady state and have the same growth rate at steady state. It is only in this case that poor regions grow faster than rich ones and eventually catch them up in the long run.

When cross-sectional data are available for two periods, initial period 0 and final period T , then Barro and Sala-i-Martin (1995) show that the hypothesis of this assumption of unconditional β -convergence is usually tested using the following model:

$$\frac{1}{T} \ln\left(\frac{y_{iT}}{y_{i0}}\right) = \alpha + \beta \ln(y_{i0}) + u_i \quad u_i \rightarrow iid(0, \sigma_u^2) \quad (16.15)$$

where $i = 1, \dots, N$, N is the number of regions in the sample; y_{iT} is the per capita output (measured, for instance, by income or per capita GDP) for region i at time t , $t = 0$ or T ; $(1/T) \cdot \log(y_{iT}/y_{i0})$ is the average growth rate of per capita output between the two dates; and α and β are the unknown parameters to be estimated. There is unconditional β -convergence if β is negative and significant. The rate of convergence between regions can then be estimated as $\gamma = -\ln(1 + T\beta)/T$.

Consider now the case of regions with different steady states. Then, as we showed before, the growth rate of a region is positively related to the distance that separates it from its own steady state. This is the concept of conditional β -convergence. In order to test for this assumption, it is necessary to hold constant the steady states specific to each region. This may be done by adding in Eq. (16.15) explanatory variables that control the heterogeneity of the long-term path:

$$\frac{1}{T} \ln\left(\frac{y_{iT}}{y_{i0}}\right) = \alpha + \beta \ln(y_{i0}) + \gamma X_i + u_i \quad u_i \rightarrow iid(0, \sigma_u^2) \quad (16.16)$$

Where X_i is the vector of variables adjusting for the steady state of region i . As before, there is conditional β -convergence if β is negative and significant. The additional variables can be divided in two groups. On the one hand, state variables in accordance with the Solow-Swan model or some version of it must be introduced. As in Eq. (16.11), these are physical capital, human capital, and population growth rate. On the other hand, empirical studies often include numerous control variables, the expected effects of which correspond to their influence on the position of the steady state. For instance, Durlauf et al. (2005) identify 145 potential

growth determinants. This concept of convergence is compatible with a high degree of inequality if the regional steady states are very different. The question then is why the steady states of some regions remain so low.

16.3.2 Space and Growth

While lots of papers analyzing convergence at subnational scales initially employed techniques used in cross-country analysis, there is recognition that countries and regions are not interchangeable. Indeed, regions usually display a greater deal of openness, and various forms of regional interdependencies exist. Consequently, a vast strand of the regional science literature has made use of spatial econometric techniques and specifications to analyze regional convergence. We briefly review here some of the main issues at stake.

One major issue associated with the spatial dimension of the data is spatial autocorrelation in the error terms. Indeed, in the cross-sectional context, units are spatially organized and the *iid* assumption usually imposed in convergence specifications is overly restrictive. Various specifications are appropriate to control for spatial dependence; we present here the most commonly used. Consider Eq. (16.16) in matrix form such as $y = X\gamma + u$ where y is a vector containing the observations of average regional growth rates and X is the matrix containing the observations on all explanatory variables: constant term, initial income, and all the other control variables – and u is the vector of error terms.

In the spatial lag model, a spatially lagged variable Wy is added as an additional explanatory variable:

$$y = \rho Wy + X\gamma + u \quad (16.17)$$

where W is the spatial weight matrix and ρ is the spatial autoregressive parameter. The error term u is *iid*. The spatial lag Wy is always endogenous so that this specification should be estimated using maximum likelihood or instrumental variables. Particular attention should be given to the interpretation of the coefficients in this model as they only include the direct marginal effects of an increase in the associated explanatory variables, excluding all indirect induced effects (LeSage and Pace 2009 and ▶ Chap. 77, “Interpreting Spatial Econometric Models” in this handbook).

The spatial error model is a special case of a nonspherical error covariance matrix in which the spatial error process is based on a parametric relation between a location and its neighbors. In the spatial autoregressive specification, the error vector u takes the form

$$u = \lambda Wu + \varepsilon \quad (16.18)$$

where ε is *iid* and λ is the spatial autoregressive parameter. Conversely, the moving average specification can be expressed as

$$u = \gamma W\varepsilon + \varepsilon \quad (16.19)$$

Both models can be estimated using maximum likelihood or generalized method of moments. The two specifications differ in the terms of the range of spatial dependence in the variance-covariance matrix and of the diffusion process they imply. In particular, in the first case, the spillovers are global: a random shock in one observation impacts upon the income of all the regions in the sample. In the second case, the spillovers remain local: a shock in location i only affects the regions directly interacting with i , that is, the regions j for which $w_{ij} \neq 0$.

In the convergence context, both models have been extensively used to capture regional interdependence (Rey and Le Gallo 2009). Interestingly, some cross-country studies also acknowledge the need of taking spatial dependence into account and hence use spatial econometric techniques. Models incorporating spatial lags of the dependent and independent variables (spatial Durbin model) or higher-order spatial models have also been suggested (for a recent review, see Fischer and Wang 2011). As the spatial Durbin model encompasses the spatial lag and the spatial error model, it can be used as a basis for specification search (see ► Chap. 27, “**Classical Contributions: Von Thünen, Weber, Christaller, Lösch**” in this handbook for more details on specification search in cross-sectional spatial models).

Finally, note that a recent trend of the literature consists in providing sound theoretical foundations for the inclusion of spatial dependence in β -convergence models. For instance, Ertur and Koch (2007) show how a spatial Durbin model version of the β -convergence model can be obtained from a theoretical growth model with Arrow-Romer externalities and spatial externalities that imply inter-economy technology interdependence. Likewise Fingleton and Lopez-Bazo (2006) introduce substantive spatial externalities in the neoclassical convergence equation and show how this leads to a different steady-state level of output per unit of effective labor than would otherwise occur.

16.3.3 Econometric Issues

Although the conditional β -convergence approach has given rise to hundreds of studies, it has also been widely criticized.

16.3.3.1 Endogeneity of Explanatory Variables

In a regression setup, error terms are often correlated with the explanatory variables, leading to endogeneity and inconsistent estimates. In β -convergence models, there are numerous sources of endogeneity.

The first source of correlation between errors and some explanatory variables is simultaneity: some explanatory variables are not exogenous, they are determined simultaneously with growth rates, and thus they may affect growth but also depend on growth. For instance, given the Solow-Swan framework, state variables such as investment, initial per capita GDP, or human capital are equilibrium outcomes, as are regional growth rates. More generally, the causality versus the correlation issue is a prevalent one in growth econometrics. On the one hand, this implies biased estimation. On the other hand, this calls into question the interpretation of regression results and the extent to which these variables affect the steady-state levels.

Finding appropriate instruments, that is, variables that are correlated with the endogenous explanatory variables but uncorrelated with, or orthogonal to, the error terms, is a difficult task. Indeed, appropriate instrumental variables are rarely available. Since growth can be explained by numerous determinants, it is difficult to identify instruments that are correlated with the endogenous variables and yet can legitimately be eliminated from the regression. Moreover, as the effect of some variables on growth may be delayed, using lagged explanatory variables as their exogenous instruments is not optimal either.

The second source of correlation between errors and explanatory variables is measurement errors or errors in variables. This is of particular concern in growth regressions. Indeed, many countries build databases in which the accuracy of the variables is undoubtedly measured with error, and also in many cases, pragmatic decisions have to be made to use a variable that is only a proxy of a true variable. When the initial per capita GDP is mismeasured, the attenuation bias tends to bias the estimates of β in favor of the β -convergence hypothesis. For instance, Temple (1998) argues that the famous result of conditional convergence of economies at a rate of 2 % per year could be entirely due to measurement error.

Correcting for this is not an easy task and is further complicated in the presence of spatial error autocorrelation. Indeed, Le Gallo and Fingleton (2012), using Monte Carlo simulations, show that OLS and instrumental variable estimation, which do not take into account spatial error autocorrelation, outperforms GMM-based and ML estimation. These results would indicate that measurement error plus a disturbance process involving spatial dependence is best accommodated by an estimation method that ignores spatial dependence. Clearly, the interaction between spatial autocorrelation and measurement errors, which are both easy to find in β -convergence models, should be further investigated.

The third source of correlation between errors and explanatory variables is omitted variables. In practice, it is unlikely that researchers are able to find all the variables controlling for the differences in steady states between regions. Hence, the error term in conditional β -convergence models will probably contain a number of omitted variables correlated with the included regressors, though if in the unlikely event they are orthogonal to the included regressors, then there is no problem. Trying to solve this by increasing the number of explanatory variables typically runs into the problem of simultaneity and possibly multicollinearity. Note that LeSage and Fischer (2008) have shown that the existence of omitted explanatory variables exhibiting nonzero covariance with variables included in the model yields a data-generating process for a growth regression that includes both an endogenous spatial lag and exogenous spatial lags (spatial Durbin growth model).

16.3.3.2 Robustness of Explanatory Variables

This critique relates to the choice of control variables and is linked to the lack of robustness of conditional β -convergence regression models. Indeed, the finding of conditional β -convergence and the subsequent estimation of the convergence rate is dependent upon a specific choice for the set of control variables. The lack of consensus about the most important growth determinants amplifies this problem:

if most regressors included in the empirical analysis are found to be statistically significant in some specification, it means that there are as many growth theories as the number of significant regressors and that it is impossible to distinguish between them. This is referred to as the problem of observational equivalence of competing theories, which is common in macroeconomic analysis generally.

Confronted by the variety of explanatory variables available for use in these regressions, Levine and Revelt (1992) employ extreme bound analysis, which consists of estimating the upper and the lower extreme bounds of a coefficient of a variable of interest across a range of different model specifications. The variable is considered to be robust if the coefficients at these extreme bounds are significant and if they maintain their signs and statistical significance across a diverse range of other included variables. Using this approach, they show that most variables tested turn out to be insignificant given additional control variables.

This approach has been criticized as being excessively conservative. More recently, the use of model averaging and Bayesian model averaging has been advocated in order to guide in the choice of control variables (Fernandez et al. 2001; Sala-i-Martin et al. 2004). In a spatial context, an additional source of uncertainty pertains to the choice of the spatial weights matrix. A Bayesian model averaging approach for selecting appropriate explanatory variables together with an appropriate spatial weights matrix has been suggested by LeSage and Fischer (2008). An alternative is to explain the variation in results by means of meta-analysis (Abreu et al. 2005).

16.3.4 Panel Estimation

If unmodeled region-specific unobserved effects on output levels are present, this implies a link between the error terms and initial output per capita. In order to correct for this, a number of researchers advocate convergence analysis via the use of panel data (Islam 1995). We have a choice as to how we model the individual effects: fixed effects, essentially dummy variables, one per region, or random effects, in which the individual-specific (region) effect is captured as a random variable. The setup of fixed effects models follows on naturally from the pure cross-sectional growth models considered thus far, typically having the form

$$\begin{aligned} \ln y_{it} &= \gamma_t + \alpha \ln(y_{it-\tau}) + X'_{it}\beta + a_i + u_{it} \quad t = 2, \dots, T \\ u_{it} &\rightarrow iid(0, \sigma_u^2) \end{aligned} \tag{16.20}$$

which can be written as a growth equation as follows:

$$\begin{aligned} \Delta \ln y_{it} &= \gamma_t + (\alpha - 1) \ln(y_{it-\tau}) + X'_{it}\beta + a_i + u_{it} \quad t = 2, \dots, T \\ \Delta \ln y_{it} &= \ln y_{it} - \ln y_{it-\tau} \\ u_{it} &\rightarrow iid(0, \sigma_u^2) \end{aligned} \tag{16.21}$$

where growth $\Delta \ln y_{it}$ is measured between period t and some previous period $t - \tau$ (usually $\tau \geq 5$ years to avoid business cycle effects). In this approach, all the unobserved time-invariant regional heterogeneity is captured by individual-specific effects, denoted by a_i . Following Eq. (16.11), the matrix X includes other possibly time-varying factors affecting growth. In addition, growth depends on the start-of-period level $\ln(y_{it-\tau})$, so the estimate of the coefficient α gives the rate of convergence. The term γ_t represents time (dummy variable) effects that are constant across locations.

The presence of the lagged dependent variable together with the time-invariant effect a_i in Eq. (16.20) renders OLS inconsistent even when the transient disturbances u_{it} are not serially correlated. The most obvious way to fix this is to first difference the data, so that the individual-specific (fixed or random) effects are eliminated. Thus, our differenced specification is

$$\Delta \ln y_{it} = \Delta \gamma_t + \alpha \Delta \ln(y_{it-\tau}) + \Delta X'_{it} \beta + \Delta u_{it} \quad t = 3, \dots, T \quad (16.22)$$

While the convergence parameter α is identified in Eq. (16.20), eliminating the time-invariant individual-specific effects does not solve the problem of inconsistent and biased parameter estimation via OLS because the lagged dependent variable is correlated with u_{it} , and there is also potential endogeneity of other regressors (including measurement error), omitted variables and spatial dependence. Rather, the reason to first difference is to create instruments that are not correlated with the individual effects.

With regard to spatial dependence, this can exist as a result of direct autoregressive interaction across space of the dependent variable, as a consequence of a spatial error process, or both. A good, comprehensive summary for static spatial panel models is provided in Chap. 12 of Pirotte (2011). If we add a spatially lagged dependent variable to the difference equation we obtain:

$$\Delta \ln y_{it} = \Delta \gamma_t + \alpha \Delta \ln(y_{it-\tau}) + \rho \Delta W_N \ln y_{it} + \Delta X'_{it} \beta + \Delta u_{it} \quad (16.23)$$

The variable $\Delta W_N \ln y_{it}$ is also endogenous, as in the pure cross-section case. While difference-GMM estimation may appear to be appropriate, by using lagged levels of variables as instruments, it does typically create a weak instrument problem. One estimator that can potentially deal with these problems is the system GMM estimator (Arellano and Bond 1991; Bond et al. 2001); this estimates Eq. (16.23) combining both the difference equation and the corresponding levels equation, with lagged first differences as instruments for the levels equation, and lagged levels for the equation in first differences. One should however use this cautiously because, using all available lags of variables as instruments, this estimator in particular presents significant practical problems relating to overfitting and thus failure to purge endogeneity. The solution seems to restrict the number of lags employed as internally generated instruments so as to clearly satisfy the relevant diagnostics, but one may still have use external instruments in order to obtain the

necessary instrument orthogonality for consistent estimation. For the additional moments conditions associated with the levels equation to be orthogonal, it is sufficient for the variables to be mean stationary, having controlled for common time effects γ_t .

The other form of spatial dependence in panel models involves the disturbances. Pirotte (2011) classifies static spatial panel models according to whether the spatial disturbance process is autoregressive (SAR), or a moving average process (SMA), and whether the individual effects are considered to be fixed (deterministic or FE), or random effects (RE). If the random individual effects are not spatially autocorrelated, but the transient component of the compound error is, then he refers to the model as RE-SAR or RE-SMA. If the spatial error process applies in the same way to both transient and individual error components, so that the spatial process is at the level of the compound errors and not its individual components, then this is referred to as SAR-RE or SMA-RE, according to whether we are considering an autoregressive or moving average specification. If however the individual effects are fixed, and spatial effects are restricted to the transient errors, then the model is referred to as FE-SAR or FE-SMA according to whether we have an autoregressive or moving average process.

Accordingly, introducing the RE-SAR (or the FE-SAR) specification to our levels model gives

$$\begin{aligned}\ln y_{it} &= \gamma_t + \alpha \ln(y_{it-\tau}) + X'_{it}\beta + a_i + u_{it} \\ u_{it} &= \lambda M_N u_{it} + \xi_{it}\end{aligned}\tag{16.24}$$

where M_N is an $(N \times N)$ matrix specific to time t , where N is the number of regions (and therefore M_N has similar properties to W_N) and a_i are random (or fixed) effects. The two forms of interactions can also be combined and one might even extend the spatial dependence in the error to include both the transient errors and the individual effects to give the spatial autoregressive equivalent of SAR-RE:

$$\begin{aligned}\ln y_{it} &= \gamma_t + \alpha \ln(y_{it-\tau}) + \rho W_N \ln y_{it} + X'_{it}\beta + \psi_{it} \\ \psi_{it} &= a_i + u_{it} \\ \psi_{it} &= \lambda M_N \psi_{it} + \xi_{it} \\ \xi_{it} &\sim iid(0, \sigma_\xi^2)\end{aligned}\tag{16.25}$$

Alternatively, the equivalent of SMA-RE entails the moving average error process involving both individual and transient errors (Fingleton 2008) with $\psi_{it} = \lambda M_N \xi_{it} + \xi_{it}$.

With spatially dependent (moving average or autoregressive) errors combined with an endogenous spatially autoregressive spatial lag, the GMM approach typically has several stages, first one uses instrumental variables, assuming no spatial error process, to obtain consistent estimates of the residuals. These then become the basis for GMM estimates of the error process parameters. Finally, the data are

purged of the error dependence and consistent estimates obtained via instrumental variables in the final stage. Overall, with these more complex models, it is evident that methods based on GMM are the most versatile because they can handle multiple endogeneity and are robust to alternative error distributions, issues that are problematic under maximum likelihood.

16.3.5 Multiple Regimes and Convergence Clubs

As we have shown above, $\beta < 0$ is consistent with the assumptions of the neoclassical growth model. However, this condition is also potentially consistent with economic alternatives, such as endogenous growth models or models with poverty traps. For instance, Azariadis and Drazen (1990) develop an endogenous growth model characterized by the possibility of multiple, locally stable steady-state equilibria. Which of these different equilibria a region will be converging to depends on the range to which its initial conditions belong? In other words there are convergence clubs, that is, groups of economies whose initial conditions are near enough to make group members converge toward the same long-term equilibrium. From an empirical point of view, the existence of convergence clubs can be inferred from the fact that while absolute β -convergence is frequently rejected for large samples of countries and regions, it is usually accepted for more restricted samples of economies belonging to the same geographical area.

While the Arariadis-Drazen model does not exhibit convergence since different initial conditions lead to different steady states, Bernard and Durlauf (1996) show that the data generated by this model will not necessarily lead to the finding that $\beta \geq 0$. Therefore, tests for β -convergence have low power against the alternative hypothesis of multiple steady states. The problem is then to distinguish evidence of club convergence from that of conditional convergence.

From an econometric point of view, the existence of multiple equilibria is characterized by parameter heterogeneity in convergence regressions. A vast range of techniques has been used in order to detect convergence clubs. Some use a priori criteria to define club members, such as belonging to the same geographical zone or having similar initial incomes. Durlauf and Johnson (1995) use regression trees (CART algorithm) where initial income and literacy rates are used to detect the convergence clubs. In the context of regional data, a number of authors have made use of exploratory spatial data analysis (ESDA) to detect spatial regimes in the data. In particular, Moran scatter plots and Getis-Ord statistics facilitate the detection of spatial clusters of high values of regional incomes and spatial clusters of low values of regional incomes. The hypothesis of β -convergence is then tested on each group (see, for instance, Ertur et al. 2006).

At the extreme, rather than partitioning the sample into regimes based on some structural characteristics, parameter heterogeneity might also be region specific. For instance, in Eq. (16.15), region-specific parameters α_i and β_i must be estimated. While varying coefficient models might be used for that purpose (see ▶ Chap. 73, “Geographically Weighted Regression” in this handbook for

a presentation of these models), we note that for regional samples, similarities in legal and social institutions, as well as culture and language, might create spatially local uniformity in economic structures. This leads to situations where convergence rates are similar for regions located nearby in space. In order to capture this combination of parameter heterogeneity and local similarity, spatial autoregressive local estimation (SALE) model has been suggested by Pace and LeSage (2004).

16.4 Sigma-Convergence and Distribution Approach to Convergence

We now turn to alternative concepts of convergence that have been used in the literature on regional growth.

16.4.1 σ -Convergence

In this approach, convergence is linked to the study of the dynamic evolution of some indicator of dispersion of output per capita between regions. The focus is then on whether this indicator increases or decreases over time. Two indicators of cross-sectional dispersion are commonly used: the standard deviation of log income or the coefficient of variation coefficient of this distribution.

Specifically, the test of σ -convergence consists of comparing an indicator of dispersion, computed at the end of the period, to the value of this indicator computed at the beginning of the period. There is σ -convergence if this indicator decreases over time. Formal tests using regression specifications have also been suggested by Carre and Klomp (1997) and Egger and Pfaffermayr (2009).

It is possible to show that β -convergence is a necessary but not a sufficient condition to σ -convergence. The point of departure is the absolute β -convergence equation where the dependent variable is the cumulated growth rate:

$$\ln(y_{iT}/y_{i0}) = a + \beta \ln(y_{i0}) + u_i \quad (16.26)$$

This equation is rewritten as

$$\ln(y_{iT}) = a + (1 + \beta) \ln(y_{i0}) + u_i \quad (16.27)$$

By taking the variance of each term in this equation, we have $V[\ln(y_{iT})] = (1 + \beta)^2 V[\ln(y_{i0})] + V(u_i)$, from which it is easy to show that

$$VR = \frac{V[\ln(y_{iT})]}{V[\ln(y_{i0})]} = \frac{(1 + \beta)^2}{R^2} \quad (16.28)$$

where R^2 is the multiple correlation coefficient associated with Eq. (16.27).

From this, it is evident that β -convergence ($\beta < 0$) is a necessary but not a sufficient condition for σ -convergence ($VR < 1$). In fact, the final result depends upon two opposite effects. The first is the existence of β -convergence implying mean reversion. The second is linked to the existence of specific shocks to which the regions are submitted and that permanently generate per capita output dispersion. σ -convergence is the result of these two mechanisms and exists if the beneficial effects of mean reversion dominate the negative effects of perturbations affecting the regions.

This concept has been subject to a number of criticisms, the first of which obviously concerns the dependence of σ -convergence on the initial date. Second, it only focuses on the second moment of the distribution and is not informative about other moments that may be of interest, such as skewness or kurtosis. Third, interpreting measures of dispersion is not straightforward when distributions are not unimodal, and it is often the case that we encounter multimodality and twin-peakedness in practice. Fourth, it is subject to a spatial identification problem. Indeed, given a map of N incomes with a sample variance σ^2 then there are $N!$ spatial permutations on the map that would have the same sample variance.

Finally, Quah (1993) forcefully argues that it does not provide meaningful information about income dynamics nor about the mobility of regions within a distribution. For instance, if two regions exchange their relative position between the initial and final date while the gap between the two remains unchanged, then the standard error of this distribution is constant over the period even if the situation of the two regions has changed radically.

16.4.2 Studying the Evolution of the Cross-Sectional Distributions

In the light of these criticisms, Quah (1993, 1996) argues that the cross-sectional distributions of income should be considered in their entirety rather than just computing one synthetic indicator such as dispersion. Indeed, with regard to σ -convergence, it tells us nothing about distribution dynamics. Rather, the evaluation of distribution dynamics can be accomplished on the basis of two criteria: the study of the evolution of income level distributions and analysis of the position of the regions or groups of regions within distributions.

Concerning the first point, the method consists of comparing the cross-sectional distributions of regional income at different points in time and evaluating the degree to which the location and shape of these distributions changes.

One possibility is to estimate, using nonparametric smoothing methods, such as kernel estimates, the density function of income for the sample, and examining the changes in the form of this density. For instance, Fig. 16.8 represents two possible ways in which the distribution might evolve over time, each representing two types of convergence. If, given the initial distribution, the regions in the sample evolve toward a tighter distribution, then there is global convergence of all regions toward the same level of income. On the contrary, if the distribution becomes bimodal or multimodal, then the regions converge toward different levels, which is symptomatic of different

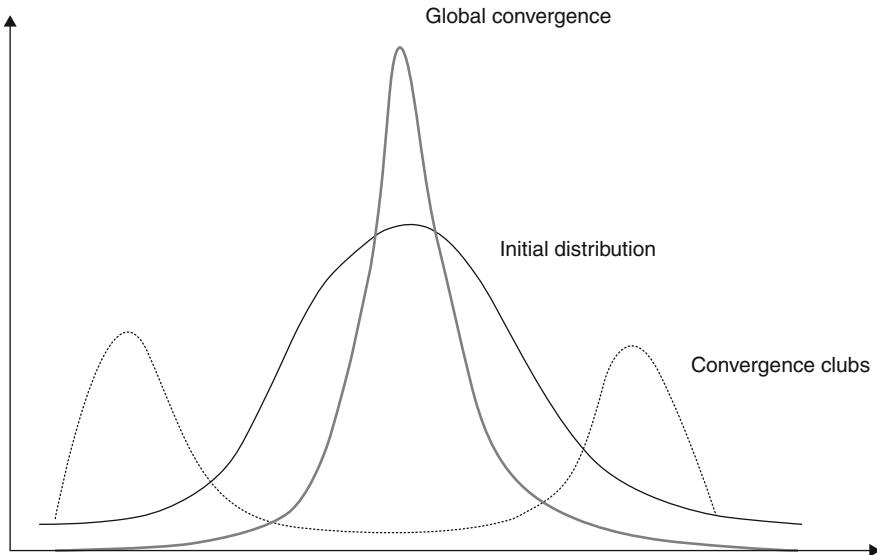


Fig. 16.8 Density functions for three different convergence issues

convergence clubs. In order to go beyond simple visual impression, tests for multimodality can also be undertaken (Henderson et al. 2008).

It is also possible to estimate cross-sectional distribution densities using mixture models, which are weighted sums of component distributions. In this case, one can say that convergence occurs when the distributions are better approximated over time by a small number of components, while multiple components are an indication of multiple regional steady states. The number of components can be evaluated using a bootstrap LR test.

Concerning the second point on the position of the regions or groups of regions within the distributions, we observe that shape dynamics does not directly address this issue. Nonetheless, it may be of interest to study whether, for a given time period, the regions have changed their relative position in the income distribution, that is, which regions move up and down in this distribution.

One method that allows detection of the movements of the regions from one period to another consists of estimating transition matrices or Markov chains. These are constructed using a discretization of the distribution of income into several classes (using for instance quartiles or quintiles of the distribution). Transition matrices allow one to estimate the probabilities of passage from one income class to another, or of remaining in the same income class, over time. If the probabilities of passage from one class to another are high, then mobility is high. If the probability of staying in the same class is high, then mobility is low. By extension, it is possible to detect whether the level of income is tending toward homogeneity or, on the contrary, if distinct groupings of regions with different incomes are emerging and being maintained over time. Formal mobility indices may also be computed, while the ergodic distribution,

that is the long-term distribution, allows one to see the type of convergence mechanism that is at work. Concentration of the frequencies in the median class would imply convergence to the mean, while concentration of the frequencies in several of the classes, that is, a multimodal limit distribution, may be interpreted as a tendency toward stratification into different convergence clubs.

In order to operationalize this, some strong assumptions are usually made, such as stationarity of transition probabilities and a first-order process. Formally, denote F_t as the cross-sectional distribution of income at time t relative to the sample average. A set of K different GDP classes is defined. If the frequency of the distribution follows a first-order stationary Markov process, then the $(K \times 1)$ vector F_t , indicating the frequency of the regions in each class at time t , is described by the following equation:

$$F_{t+1} = MF_t \quad (16.29)$$

where M is the $(K \times K)$ transition probability matrix representing the transition between the two distributions. If the transition probabilities are stationary, that is, if the probabilities between two classes are time invariant, then

$$F_{t+s} = M^s F_t \quad (16.30)$$

The ergodic distribution of F_t is approached as s tends toward infinity in Eq. (16.30). Such a distribution exists if the Markov chain is regular, *that is*, if and only if for some N , M^N has no zero entries. In this case, the transition probability matrix converges to a limiting matrix M^* of rank 1. The existence of an ergodic distribution, F^* is then characterized by

$$F^*M = F^* \quad (16.31)$$

Each row of M^t tends to the limit distribution as $t \rightarrow \infty$. According to Eq. (16.31), this limit distribution is therefore given by the eigenvector associated with the unit eigenvalue of M . The estimation of the transition matrix is based on maximum likelihood estimation.

As indicated, strong assumptions must usually be made to estimate such transition matrices. Moreover, the results are sensitive to the number and size of the groups of observations used to discretize the data. In fact, discretization of the state space may significantly alter the probabilistic properties of the data.

To overcome this problem of sensitivity of the results to the discretization, stochastic kernels have been suggested. They are the continuous counterpart of transition probability matrices. Formally, if $f_{X(t)}$ is the regional income density for n regions in period t , then the evolution of the cross-sectional distribution is modeled as

$$f_{X(t+s)} = \int_{-\infty}^{\infty} M_{t,s} f_{X(t)} dx \quad (16.32)$$

where $M_{t,s}$ is the stochastic kernel representing where points in $f_{X(t)}$ move to in $f_{X(t+s)}$. The estimation of this kernel may be based on an estimate of the conditional distribution. In order to explore the transitional dynamics provided by this approach, three-dimensional representations and two-dimensional contour plots are used. For example, polarization or convergence clubs in the per capita GDP distribution are reflected in peaks in the 3D kernel or by concentrated values in the contour plot. Fischer and Stumpner (2008) introduce three-dimensional stacked conditional density plots and highest density regions plot for the visualization of the transition function.

16.4.3 Distribution Dynamics and Space

As in the confirmatory econometric analysis of growth and convergence, the spatial dimension of the data invalidates some of the restrictive assumptions regarding random sampling on which σ -convergence and distribution dynamics rest. We briefly consider in this section how this impacts on the measures of convergence and distribution dynamics. First, we note that the concepts of convergence that have been developed in the preceding sections must be adjusted to take into account spatial autocorrelation in the data. Secondly, we observe that much work has been done in exploratory spatial data analysis (ESDA) and Exploratory Space-Time Data Analysis (ESTDA), and their application to convergence and growth analysis has led to interesting new insights.

Regarding the first point, consider the σ -convergence measure presented earlier. We have already pointed out that it is uninformative with regard to the morphology of the distribution and the degree of intradistributional mobility. Moreover, in a spatial context, the presence of spatial dependence complicates the interpretation of, and inference based on, this concept. For instance, Rey and Dev (2006) show that the sample variance also reflects the level and structure of spatial dependence in the data. This should be purged in order to correctly interpret this concept of convergence.

Similarly, spatial autocorrelation has been incorporated into measures of intradistributional dynamics. In the case of discrete Markov chains, Rey (2001) extends the approach by estimating transition matrices subject to the spatial lag of the income values for each region. This allows one to analyze how the spatial environment affects the transition probabilities of a region through the income distribution. It is usually found that the probabilities of a given region staying in the same class or of moving up one class are ameliorated when the region is surrounded by other wealthy regions.

Spatial autocorrelation must also be considered when analyzing the shapes of per capita GDP distributions and when estimating stochastic kernels. This is done using regional conditioning, that is, basing density function and kernel estimation on a region's income expressed relative to its geographical neighbors. A formal inferential framework to test hypotheses about distribution dynamics in the presence of spatial effects still needs to be developed however.

Regarding the second point, traditional convergence measures can be usefully augmented by the different ESDA and ESTDA measures. First, the classical Moran's I statistics is naturally used to assess the level of spatial dependence in the income series and its evolution over time.

Second, local measures of spatial autocorrelation can also be used. In particular, local spatial instability is studied by means of the Moran scatterplot, which plots the spatial lag of standardized income against the original values. The four different quadrants of the scatterplot correspond to the four types of local spatial association between a region and its neighbors: HH denotes a region with a high value surrounded by other regions with high values and LH indicates a low value region that is surrounded by regions with high values, etc. Quadrants HH and LL (resp. LH and HL) refer to positive (resp. negative) spatial autocorrelation indicating local spatial clustering of similar (resp. dissimilar) values. This approach has been used extensively to analyze the evolution of the spatial distribution of income in several regional samples. Whenever these Moran scatterplots are constructed for several years, Rey (2001) has suggested using the discrete Markov methodology: in any period, there are four possible states: HH, LL, LH, and HL so that between any two periods, 16 different spatial transitions are possible, which can be summarized by a transition probability matrix.

16.5 Conclusions

We have reviewed alternative approaches to regional growth and convergence empirics, focusing on the various methodological problems and solutions that have been offered in the literature. Clearly, there is no obvious consensus regarding the most appropriate approach, or modeling strategy, or even whether convergence is actually a real phenomenon, or simply a feature of the theoretical model that has been the dominant one in the literature, the neoclassical growth model.

There is a point at which some of these approaches do give comparable conclusions, however. Indeed, we can obtain estimates of the time it will take for economies to converge both from the neoclassical growth model and from the Markov chain approach. According to Fingleton (1999), for the regions of the EU, the time needed to achieve neoclassical (conditional) convergence will be of the order of 200–300 years. This is simply due to diminishing returns to capital setting in very slowly, that is, effectively α is close to 1 in the model of Sect. 16.2. Under the Markov model, convergence to the ergodic distribution should, it is estimated, take a similar amount of time, at least 300 years. Of course, the latter is stochastic convergence, implying constant probabilities of different income states but allowing movement of regions across income states. It is evident that, for the EU at least, convergence of some sort, if it occurs at all, will not be a rapid phenomenon and be characterized by distributed income levels rather than the homogeneity associated with unconditional neoclassical convergence.

References

- Abreu M, De Groot HLF, Florax RJGM (2005) A meta-analysis of β -convergence: the legendary 2 %. *J Econ Surv* 19:389–420
- Arellano M, Bond S (1991) Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations. *Rev Econ Stud* 58:277–297
- Azariadis C, Drazen A (1990) Threshold externalities in economic development. *Q J Econ* 105:501–526
- Barro RJ (1991) Economic growth in a cross section of countries. *Q J Econ* 106:407–443
- Barro RJ, Sala-i-Martin X (1995) Economic growth theory. McGraw-Hill, Boston
- Bernard A, Durlauf S (1996) Interpreting tests of the convergence hypothesis. *J Econ* 71:161–173
- Bond S, Hoeffler A, Temple J (2001) GMM estimation of empirical growth models. Centre for Economic Policy Research Discussion Paper No. 3048
- Carree M, Klomp L (1997) Testing the convergence hypothesis: a comment. *Rev Econ Stat* 79:683–686
- Durlauf S, Jonhson P (1995) Multiple regimes and cross-country growth behavior. *J Appl Econ* 10:365–384
- Durlauf S, Johnson P, Temple J (2005) Growth econometrics. In: Aghion P, Durlauf S (eds) Handbook of economic growth. North-Holland, Amsterdam
- Egger P, Pfaffermayr M (2009) On testing conditional sigma-convergence. *Oxf Bull Econ Stat* 71:453–473
- Ertur C, Koch W (2007) Growth, technological interdependence and spatial externalities: theory and evidence. *J Appl Econom* 22:1023–1062
- Ertur C, Le Gallo J, Baumont C (2006) The European regional convergence process, 1980–1995: do spatial dependence and spatial heterogeneity matter? *Int Reg Sci Rev* 29:2–34
- Fernandez CE, Ley E, Steel M (2001) Model uncertainty in cross-country growth regressions. *J Appl Econ* 16:563–576
- Fingleton B (1999) Estimates of time to economic convergence: an analysis of regions of the European Union. *Int Reg Sci Rev* 22:5–35
- Fingleton B (2008) A generalized method of moments estimator for a spatial panel model with an endogenous spatial lag and spatial moving average errors. *Spat Econ Anal* 3:27–44
- Fingleton B, Lopez-Bazo E (2006) Empirical growth models with spatial effects. *Pap Reg Sci* 85:177–198
- Fischer MM, Stumpner P (2008) Income distribution dynamics and cross-region convergence in Europe. Spatial filtering and novel stochastic kernel representations. *J Geogr Syst* 10:109–139
- Fischer MM, Wang J (2011) Spatial data analysis: models, methods and techniques. Springer, Heidelberg/New York
- Henderson D, Parmeter C, Russell R (2008) Modes, weighted modes, and calibrated modes: evidence of clustering using modality tests. *J Appl Econ* 5:607–638
- Islam N (1995) Growth empirics: a panel data approach. *Q J Econ* 110:1127–1170
- Le Gallo J, Fingleton B (2012) Measurement errors in a spatial context. *Reg Sci Urban Econ* 42:114–125
- LeSage JP, Fischer MM (2008) Spatial growth regressions: model specification, estimation and interpretation. *Spat Econ Anal* 3:275–304
- LeSage J, Pace K (2009) Introduction to spatial econometrics. CRC Press, Boca Raton
- Levine R, Revelt D (1992) A sensitivity analysis of cross-country growth regressions. *Am Econ Rev* 82:942–963
- Mankiw NG, Romer D, Weil DN (1992) A contribution to the empirics of economic growth. *Q J Econ* 107:407–437
- Nijkamp P, Poot J (1998) Spatial perspectives on new theories of economic growth. *Ann Reg Sci* 32:7–37
- Pace R, LeSage J (2004) Spatial autoregressive local estimation. In: Getis A, Mur J, Zoller H (eds) Spatial econometrics and spatial statistics. Palgrave Macmillan, Basingstoke

- Pirotte A (2011) Econometrics of panel data. *Economica*, Paris
- Quah D (1993) Empirical cross-section dynamics in economic growth. *Eur Econ Rev* 37:426–434
- Quah D (1996) Empirics for economic growth and convergence. *Eur Econ Rev* 40:1353–1375
- Rey SJ (2001) Spatial empirics for economic growth and convergence. *Geogr Anal* 33:195–214
- Rey SJ, Dev B (2006) σ -convergence in the presence of spatial effects. *Pap Reg Sci* 85:217–234
- Rey SJ, Le Gallo J (2009) Spatial analysis of economic convergence. In: Mills TC, Patterson K (eds) *Palgrave handbook of econometrics*, vol 2, Applied econometrics. Palgrave MacMillan, Basingstoke
- Sala-i-Martin X, Doppelhofer G, Miller R (2004) Determinants of long-term growth: a Bayesian averaging of classical estimates (BACE) approach. *Am Econ Rev* 94:813–835
- Temple J (1998) Robustness tests of the augmented Solow model. *J Appl Econ* 13:361–375

Charlotta Mellander and Richard Florida

Contents

17.1	Introduction	318
17.2	The Firm/Industry Focus in Regional Research	319
17.3	Postindustrialism and the Knowledge Economy	320
17.4	Human Capital	321
17.5	Occupations and the Creative Class	322
17.6	The Organizing Role of Place	325
17.7	Conclusions	327
	References	328

Abstract

The past couple of decades have seen what amounts to an intellectual revolution in urban and regional economic research concerning the role of skills in economic growth. From industrial location theory and Alfred Marshall's concern for agglomeration to more recent research on high-tech districts and industrial clusters, firms and industries have been traditionally the dominant unit of analysis. But since the 1990s, there has been a growing focus on skills. This broad research thrust includes studies of human capital; the creative class and occupational class more broadly; and physical, cognitive, and social skills, among others. This research highlights the growing geographic divergence of skills across cities and metros and their effects on regional innovation, wages, incomes, and development broadly. An expanding literature notes the growing

C. Mellander (✉)

Jönköping International Business School, Jönköping, Sweden
e-mail: Charlotta.Mellander@jibs.hj.se

R. Florida

Rotman School of Management, University of Toronto, Toronto, ON, Canada
e-mail: richard.florida@rotman.utoronto.ca

importance of place in organizing and mobilizing these skills. Studies have focused on the role of amenities, universities, diversity, and other place-related factors in accounting for the growing divergence of skills across locations. This chapter summarizes the key lines of research that constitute the skills revolution in urban and regional research.

17.1 Introduction

A large and influential body of research on the role of skills in regional economic growth and development has emerged over the past decade or so. This work on skills has greatly expanded our understanding of the nature, role, and mechanics of regional growth and development. This relatively recent line of research can be divided into two main thrusts – studies that focus on and measure human capital in terms of educational attainment and those that measure skills via occupations.

This work represents something of a break with an earlier focus on the role played by firms and industries. Classical location theories, such as developed by Weber, von Thünen, and Christaller before World War II (McCann 2001), emphasize the transportation costs and tradeoffs made by large industrial firms. Marshall long ago noted forces and factors that cause firms to agglomerate. Building on his work, economists have catalogued the factors including proximity to natural resources and transportation routes, shared inputs, knowledge spillovers, and access to labor that lie behind such agglomerations or clusters. During the 1980s, economic geographers identified the rise of industrial districts and flexibly specialized industrial networks as alternatives to vertically integrated production by large firms (Piore and Sabel 1984). Porter (1990) identified four key reasons behind geographic clusters including factor conditions, demand conditions, related and supporting industries, and strategy, structure, and rivalry. Others have noted the role of social factors such as social capital and trust as additional factors orienting geographic clusters (e.g., Saxenian 1994).

Interest in skills and human capital dates back a long way. In his classic work on *The Wealth of Nations* (1776), Adam Smith identified the “acquired and useful abilities of all the inhabitants or members of the society” as something akin to a “fourth factor of production” operating alongside land, labor, and production, noting that “The greatest improvement in the productive powers of labour, and the greater part of the skill, dexterity, and judgment with which it is anywhere directed, or applied,” he wrote, “seem to have been the effects of the division of labour” (Smith 1776; Book 1, p. 7). Still, until recently, the great preponderance of economic and regional research has focused on the firm and firm location in order to understand geographical differences in economic performance.

Nearly a half century ago, however, Jane Jacobs (1969) noted that what distinguished cities and propelled their economic growth and development was not firms, but the geographic clustering of diverse and talented people in cities. In the mid-1980s, attention was called to the role of occupations in regional development and that we need to look beyond the products a city makes and instead

examine the roles workers play and the skills they perform. This in turn will shift the emphasis from the industry perspective to more long-run power and potential of local (human) resources (Florida 2002). Building on Jacobs, Lucas (1988) suggested that knowledge is embodied in human beings rather than in industries and that human capital externalities that stem from a concentration of highly educated individuals are the key motivating force in economic growth and development. Over the past decade or so, a growing body of research has identified the growing concentration and geographic clustering of skill as the key factor in regional growth and development. This research has noted a growing divergence in the geographic location and concentration of skills and their increasing importance to regional innovation, productivity, and growth. A debate has emerged over how best to measure skill. There is a long literature on the role of educational-based human capital, at the regional as well as national levels, in economic performances. More recent research has shifted attention to occupation-based measures of skill and also to more discrete measures of occupational skill itself.

Some have also invoked the classic chicken and egg question: what comes first they ask – firms or skills? In our view, this is a misleading question. It can be better framed in terms of demand and supply, where firms comprise the demand for skill, with skills thus functioning as supply. This is not so much a chicken and egg or either/or question, but a question of how firms and skills interact to inform regional economic growth. The analog is of course producers and consumers: one cannot entangle which of these came first. They require each other.

This chapter focuses on the growing attention paid to, and the importance of, skills for regional economic development and growth. It summarizes the major thrusts and contributions of the field. In doing so, it distinguishes between education-based and occupational definitions of skills. It discusses the opportunity to link firms and skills – or industries and occupations – in order to gain more robust theories and empirical research on the key factors that shape regional economic growth and development.

17.2 The Firm/Industry Focus in Regional Research

Historically, regional economists, such as von Thünen, Weber, Christaller, Ohlin, Hoover, Lösch, and others, have focused on the firm as the unit of analysis (McCann 2001). This also stems from a strong tradition built on the work by Marshall in the 1890s, who argued that firms cluster to achieve the advantages of collocation, such as shared labor markets, shared inputs, risk minimization, and knowledge spillovers. Hotelling a few decades later showed how firms producing similar types of products and that do not compete based on price have incentives to collocate next to one another. This focus on the firm developed just as the economy was moving away from agriculture toward large industrial firms and heavy industry. This was a time when transportation costs were high and location decisions were heavily influenced by proximity to natural resources (McCann 2001).

This was the backdrop for the focus on the firm and, in particular, on the large firm as the dominant unit of analysis in urban economics, economic geography,

and regional science. Many studies focused on the location and site selection decisions by these firms and later on the location choices of their branch plants. Vernon (1966) advanced a simple model of industrial location based on the product cycle – firms would decentralize production through branch plants once production processes became standardized. Others later focused on the growing international spatial division of labor informed by the location decisions and global reach of these multinational firms (e.g., Massey 1984). There was little inclination to examine the role of human capital or occupations as most jobs were standardized. When labor was considered, it was mainly seen as a cost to minimize.

17.3 Postindustrialism and the Knowledge Economy

This changed however with the shift from an industrial to postindustrial economy. Machlup (1962) and Drucker (1993) described the rise of the “knowledge economy” and of “knowledge workers.” Bell (1973) identified the shift to a “postindustrial society” with a new class structure based on scientists, managers, administrators, and engineers. Andersson (1985) emphasized the broader societal change from the manufacturing-based economy to the “C-society” based on creativity, communication, culture, and knowledge. Kenney and Florida identified the rise of new systems of industrial production, especially among Japanese firms, which gained competitive advantage by tapping the knowledge and intelligence of factory workers (Florida 2002).

The 1980s saw increasing interest in the shift from vertically integrated to flexibly specialized production systems (Piore and Sabel 1984). This body of research identified the rise of networked production systems based on “flexible specialization.” As production standardization decreased, workers were enabled to have a wider range of expertise, which led to a continuous firm learning process.

This distinction between knowledge-based production and more standardized goods production has several implications for regional development. First, standardized production can take place almost anywhere, (where labor and land rent are cheap), and the final product can thereafter be sent to the market place for consumption. In other words, production and consumption do not need to take place in the same location. Knowledge production, on the other hand, is most often related to service goods, where there is a need for producers and consumers to meet contemporaneously in the market place. Knowledge products are therefore in general more distance sensitive and more attached to the region where the economic activity is located. Second, knowledge and creative workers not only function as producers of high value goods, they also consume them. Glaeser et al. (2001) describe how increased average incomes based on the reallocation of labor into more productive sectors have changed the role of the regions. As incomes rise, people demand more normal and luxury goods rather than necessity goods, and those will mainly be provided in bigger cities. Higher incomes also increase the opportunity cost of not working as well as the cost for commuting. Altogether, the increasing incomes and increased cost for commuting create stronger incentives to locate in cities.

The distinction between knowledge-based and standardized manufacturing industries can be problematic when industry data is being used. The same industry codes may imply very different functions and tasks, and those often come with a spatial fragmentation. Multinational companies often locate the labor-intense low-skilled functions in less developed, low-wage countries, but keep the more high-skilled functions in the western world. To best illustrate this, one needs to understand the occupational and educational structure within the growing creative and knowledge-based industry sectors compared to the one within more traditional sectors.

17.4 Human Capital

Most research on skills is organized around the construct of education-based human capital. Human capital theory postulates that wages rise with the level of knowledge or skill (Becker 1964) and a traditional Mincer (1974) equation suggests that the hourly wage is a function of skills and education. Optimally, wage levels offered by employers are determined by the value of workers' marginal product and therefore declining when employment increases. On the other hand, human capital available in the market will be positively related to wages offered. At the aggregated level, wages are thus set by the regional supply and demand for labor. Regional wage levels would thereby be directly related to regional labor productivity. At the micro level, this may be distributed unevenly. Two regions can reach the same wage levels based on (i) a homogenous labor force or (ii) a labor force consistent of high- and low-knowledge labor that together reach the same result. But at the aggregate level, the regional wage level will reflect regional labor productivity. Wages can also be a function of a number of other factors, such as gender, immigration background, and race. Becker suggests that discrimination, for example, based on factors such as race and gender, may disturb the relation to wages, and also decrease productivity, since some people may be hired based on individual traits than how suitable they are for the job. Furthermore, wages can also be related to the available amenities, with, for example, a nice climate compensating for a lower wage, all else being equal.

A wide range of empirical studies have documented the role of human capital in national growth. A number of studies (e.g., Barro 1991; Mankiw et al. 1992) provide empirical support for this basic model. However, these studies do not account for regional spillovers and interdependencies. Fischer (2011) extends this basic approach utilizing spatial econometrics which takes technological interdependencies into account to examine the connection between knowledge diffusion and growth across regions in 22 European countries.

There is also a wide range of research on the contribution of human capital to city or metropolitan level growth (e.g., Glaeser and Saiz 2003). This research stream notes that the metro may be a more appropriate context to evaluate the effects of human capital. They note that it is harder to estimate the human capital (educational effect) on economic growth for nations than for cities, since national

growth to a larger extent will be affected by institutional differences as well. Cities also make up economically delimited units, while regional and national divisions otherwise tend to be more arbitrary and based on political decisions.

While human capital externalities affect regional productivity in cities, there are also more straightforward explanations to why we would expect a relation between higher levels of human capital and productivity. Human capital has become more unevenly distributed and concentrated over time. Berry and Glaeser (2005) have shown how human capital levels have become more and more concentrated over the last century, and how this is an endogenous self-reinforcing process where places with initially high values have increased their human capital levels more over time than other places that started off a lower position.

Ullman (1958) was one of the first to note the role of human capital for regional development. Ever since, considerable research has found significant relations between education levels and wages in cities and metropolitan areas. Rauch (1993) found that human capital intense cities are more productive and that an increase by 1 year in average education increases productivity by 3 %. Glaeser and Saiz (2003) provide empirical evidence on the correlation between human capital and regional economic growth. Firms locate in areas of high human capital concentration to gain competitive advantages, rather than letting suppliers' and customers' geography alone dictate their location. They find that skilled cities grow, relative to less skilled cities, through increases in productivity.

17.5 Occupations and the Creative Class

While the great preponderance of research on regional development uses human capital as a proxy for skill, a second approach has also emerged over the past decade which suggests that occupations may be a better proxy for skill. Thompson long ago suggested the need to utilize occupational analysis in regional development research (Florida 2002). Feser (2003) later suggested that it is more important to focus on what regions do rather than what they make in order to understand regional development and that occupations then should be the natural unit of analysis.

Florida (2002) used occupational analysis to divide the workforce into three main occupational classes – the creative class, working class, and service class. Florida based his research on Bureau of Labor Statistics data on occupations. The creative class works with knowledge, the working class is engaged in physical work, and the service class performs routine service. The creative class is divided into two subgroups: the super-creative core (computer and math occupations; architecture and engineering; life, physical, and social science; education, training, and library positions; arts and design work; and entertainment, sports, and media occupations) and the creative professionals (management occupations, business and financial operations, legal positions, healthcare practitioners, technical occupations, and high-end sales and sales management).

This approach focuses on creativity itself and not education as a proxy for skill. Research in psychology has shown that creativity is a fundamental and intrinsic

human capability. Where Marx and other classical economists looked at physical labor – in other words, the ability of humans to transform nature, create farms, and build manufactured products – in reality, it is our underlying creativity which makes us different from other species. According to Sternberg and Lubart (1999), this is what entrepreneurs, CEOs, artists, and technologists, as well as all children, have in common – our creative capacity. Creativity can be defined as the development of new ideas – as embodied in products, practices, services, or procedures – that are potentially useful. Creativity can be separated from intelligence and education, and these three factors are regarded as both substitutes and complements in the productive process carried out by individuals. Creativity can also be seen as a necessary condition to adopt and react to the constant changes around us. On an individual level, there is a close relation between creativity and productivity, and creative people are even proven to be happier in general as well as more committed and self-actualized. Personality and cognitive characteristics make some individuals more creative than others, and the level of creativity is often affected by the social and organizational context (Sternberg and Lubart 1999). Poincaré made comparisons between the creativity of a mathematician and an artist in his famous work *Mathematical Creation*. The mathematician is described as an artist, rather than a scientist. Halmos makes a comparison between the creativity in mathematics with the creativity within painting: “mathematics is never deductive in its creation” – “perhaps the closest analogy is between mathematics and painting... Almost every aspect of the life and of the art of a mathematician has its counterpart in painting.” Sternberg and Lubart (1999) also notes that creativity is a common denominator across disparate fields: “If one wanted to select the best novelist, artist, entrepreneur, or even chief executive officer, one would most likely want someone who is creative.”

Florida (2002) identifies three types of creativity: (i) technological creativity or innovation, (ii) economic creativity or entrepreneurship, and (iii) artistic or cultural creativity. He argues that these three types are mutually dependent and that they simulate and reinforce one another. Creativity also comes with costs. It is often an uncertain process that includes the risk for failure, stress, and other negative effects. Creative ideas challenge established norms and bring disorder, which imply a risk since creative people tend to be met with resistance and skepticism. It can be very difficult to change systematic beliefs. One can even argue that the only reason that science will change is because old scientists die. Similar conditions hold for, for example, arts, music, and poetry, where orthodoxy can become a constraint on novelty and individual expression. Close interaction is needed when novel ideas are introduced in order to overcome the skepticism of the audience. Not only must new ideas be presented, they need to be accepted as well (Florida 2002). Jacobs's (1969) saw cities as arenas for the generation of new ideas, and as ideal places for creative industries, whose production processes are related to higher risks, shorter life cycles, and often unique products.

The occupational approach differs from the more conventional human capital perspective in how skill is conceived and measures. The educational-based human capital is purely supply based, in other words, the amount of knowledge offered at the labor market. On the other hand, it says very little about how this knowledge is

being used. The creative class is a combination of supply and demand driven qualified labor. In order to be included in the creative class, you do not only have to have a certain amount of skill, knowledge, or creativity to offer at the labor market – the labor market must also be willing to pay for that kind of skill.

While human capital and creative occupations overlap, they are not the same thing. Research based on micro-data for the USA, Canada, Sweden, and Denmark illuminates this difference. For the USA, 35 % of the labor force belong to the creative class, and 29 % have a university degree. Out of everybody with a university degree, only 72 % have a creative job, and out of all creative class workers, only 59 % have a university degree (Florida 2012). Other research examined the differences between the occupational-based creative class and educational-based human capital for US metropolitan regions (Florida 2012) and finds that human capital is more related to regional income levels, while creative class is more linked to average wage levels, and also suggests that wages are more place based than income. Incomes can be earned and transferred more easily from other places (and even globally), while wage structures relate to the industrial and occupational setup in the region where the wage is being earned. Florida (2012) finds that the creative class significantly outperforms human capital in order to explain regional wage levels in Sweden.

McGranahan and Wojan (2007) find that the creative class is a strong explanatory factor not only in bigger cities but also in rural areas. However, they suggest that the measure becomes even more powerful if education, training, and library positions, as well as healthcare practitioners, are excluded in the measure. There is evidence for European regions that the creative class is significantly related to regional economic performance. Boschma and Fritsch find that creative class is a stronger explanatory variable than education-based human capital in relation to innovation and new firm formation, but that human capital is more strongly related to patenting (Florida 2012).

An important wave of recent studies has focused more on *skills* themselves. Rather than examining the occupation, there is instead a focus on the skills and knowledge required by that occupation. This research uses new data from the US Bureau of Labor Statistics O*NET database which collects detailed information on the actual skill content of work for more than 800 individual occupations. Bacolod et al. (2009) examined the association between four key skills – physical, motor, cognitive, and social skills – and found higher returns to cognitive and social skills, which in turn can help explain the decreased gender wage gap. They also observed a connection between skill type, finding social and cognitive skills to be associated with larger cities and metros. Feser (2003) identified the general knowledge requirements across occupations and the economic returns they generate. Feser also shows the need for an occupation-based analysis, where similarities and dissimilarities across occupations are taken into account and grouped in a more meaningful way, in order to understand regional development. Gabe (2009) differentiates between skill requirements and the returns they offer in private and public sectors. He also shows how spillover effects enhance earnings in metropolitan regions with higher shares of high-knowledge occupations.

Scott (2009) examined the connection between skills and regional employment, finding the largest increases in regional employment to be associated with cognitive-intensive occupations, with substantial employment declines for occupations that depend on physical skills. Florida (2012) defines three skill sets – analytical, social, and physical – and notes the increasing returns generated by analytical and social skills over time and decreasing returns by physical skill. They reinforce the key findings of Bacalod et al. (2009) regarding the concentration of analytical and social skills in larger metros. They conduct regression analysis including factors for both high-tech industry concentrations and education levels and find that analytical and social skills still are significantly associated with higher wage levels.

17.6 The Organizing Role of Place

While industrial production was organized in and around firms, knowledge-based or creative production is organized in cities. For decades, many assumed that advance in technology would underpin far-flung globalization enabling production to be broken apart and people to live virtually anywhere. Many proclaimed the end of geography, the death of distance, the decline of place, and the flattening of the world. But that is only half the story. As Porter (1990) has noted, many predicted the rise of globalization and end of place, and that confused many otherwise smart people. But the empirical fact is that more than half the world's population live in cities and urban areas, the highest at any point in history (Florida 2008). And the economic activity produced by the biggest metro areas account for a substantially greater economic value than their population size. The top ten metros, which house approximately 2.6 % of the world's population, account for more than 20 % of global economic activity (Florida 2008). Glaeser (2011) identifies cities as the world's key economic actors, indicating a triumph of the city. Cities not only make people and industries more productive, they even improve the conditions to the least well-off. Glaeser et al. (2001) suggest that this is the result of the relatively lower cost of transportation of the *latest* ideas and information, which tend to be more time-sensitive and in a higher demand of face-to-face interaction and urban density. This is in line with Keynes who, in *Essays in Persuasion*, emphasized the need for didactics and persuasion in relation to creative processes. Not only should new and innovative ideas be presented, they need to overcome the skepticism of the audience, which in practice only can be made through close interaction (Florida 2002).

As early as 1969, Jacobs leveled a fundamental critique of Adam Smith's notion of the division of labor captured in his famous pin factory example. Smith's story, she argued, emphasized efficiency. But the key to economic growth is innovation. Innovation, she continued, comes principally not from firms, but from cities which enable the constant combining and recombining of key inputs, including skilled people. While firms deepen and specialize the division of labor, cities, with their clustering force and combination and recombination of skilled individuals, give rise to new innovations and economic development.

Building on Jacobs' insight on cities (1969), Lucas (1988) modeled the importance of investment in education for productivity. His insights with respect to the engines of economic growth contributed to him being awarded the Nobel Prize. Lucas later said that Jane Jacobs was the one who should have received it for her contributions. Lucas also showed how collocation of skilled, talented, ambitious, and entrepreneurial people led to so-called human capital externalities. Having many driven people in the same locale will lead to spontaneous interaction and activities where they learn from one another, without any specific cost related to it. Lucas formalized the role of dense urban areas which localize human capital and information, create knowledge spillovers, and become engines of economic growth. Cities reduce the cost of knowledge transfer, so ideas move more quickly, in turn giving rise to new knowledge more quickly.

Caves (2000) showed how creative industries are more likely to be organized as geographic clusters of creative individuals as opposed to vertically integrated firms. Creative industries have higher levels of uncertainty and production challenges due to multiplicative production functions, where every input is non-substitutable and therefore must be present in order to produce. Further, shorter life cycles and a constant need for reinvention demand a closer interaction between consumers and producers, as well as more new skill combinations for a faster generation of new ideas. For these reasons, creative industries tend to be organized flexibility in places rather than in vertically integrated firms.

Florida (2002) focuses on the characteristics of places that attract highly skilled individuals. Building on research by, for example, Roback (1982) on migration across regions, Florida (2002) suggests that in addition to the human capital externalities and productivity and innovation enhancing functions of places, they also act on the consumption preferences of skilled individuals. This reasoning is a considerable advance over earlier theories that state that comparative advantage stems merely from business friendly environments, lower taxes, and lower overall firm costs. Clark et al. (2002) argue the regional winners will be those that maximize individuals' utilities and not just their incomes.

Glaeser et al. (2001) also examined the factors that attract skilled labor to cities and come to the conclusion that the most attractive cities are those that offer consumption rather than production. Cities with a diverse consumption of goods and services are more appealing than cities that do not. Glaeser et al. (2001) found that higher amenity cities attract more skilled labor and grow faster. This result is also in line with Florida (2012) who finds that individuals who find their communities aesthetically appealing are more satisfied with their places and that beauty is a stronger explanatory variable than, for example, job opportunities, housing markets, and high-quality public services. Florida (2012) also illustrates how place-specific characteristics, such as the ability to meet and make friends, quality of public schools, and being able to get from one place to another without too much traffic, are significantly more related to community satisfaction and the likelihood of staying in a place, even more so than economic variables, such as the ability to get a job or perceived future economic conditions, or individual characteristics.

Florida argues that skill needs to be considered not as a stock but as a flow of ideas that are highly mobile. A key factor here is low barriers to entry also referred to as tolerance (Florida 2002). Smart, talented people, whether in the form of human capital or creative class, are attracted to open and tolerant places where ideas will be accepted and can float freely between individuals. It will also improve the chances of a meritocracy based labor market, which is necessary for it to function efficiently, which Becker noted in the 1950s. Further, not only is openness and tolerance a condition for attracting talent, it is also a way of increasing the probability of turning new ideas into economic value.

Research by Inglehart finds that openness and tolerance are related to economic growth in studies covering more than 60 countries over four decades (Inglehart 1989). According to Inglehart, the best indicator of national tolerance is openness to gay and lesbian people. Florida (2002) used a similar variable, concentration of the same population group, to proxy for regional openness in the USA. Florida et al. (2008) later used it in a multivariate context and found it a significant contributor in order to explain the distribution of both highly educated and the creative class across metropolitan areas.

17.7 Conclusions

This handbook chapter has traced the skills revolution in urban and regional research. The long sweep of urban economics, regional science, and economic geography research has focused on the firm and industry as the key unit of analysis. But the past couple of decades have seen increasing concern for skills. The skills revolution is rooted in the changing nature of the economy – from an older industrial economy to a newer one based upon knowledge, innovation, and skill. Research has focused on the role of human capital, the creative class and occupational classes, and on skills themselves. The skills revolution has also informed increased concern for and understanding of how places operate. Research has focused on the role of amenities, universities, diversities, and other characteristics of places in accounting for the dispersion, uneven location, and geographic mobilization of skills.

We suggest it is time to get beyond the either/or focus on firms and skills, and industries and places. Both are veritable flipsides of the same analytic coin. The proverbial chick and egg problem is in our view a false dichotomy. Both firms and skills, and industries and places play a key role in and are required for regional development, a complex iterative development process. An important task for future research is to bring these foci together. With recent conceptual advances combined with the availability of micro-data on skills and firms, it is now tie to bring the two together. An important line of future research needs to identify the ways firms and skills work together to structure regional growth and development. The skill revolution has been a powerful one, advancing our understanding of the importance of skills, their growing divergence over space, and the way they power

regional growth and development as well as identifying the key aspects and functions of place. Future research promise even greater advances of how firms and skills come together to power innovation and growth. It is an exciting time to be working in this incredibly rich field.

References

- Andersson ÅE (1985) Creativity – the future of metropolitan regions. Prisma, Stockholm
- Bacolod M, Blum B, Strange W (2009) Skills in the city. *J Urban Econ* 65:136–153
- Barro RJ (1991) Economic growth in a cross section of countries. *Q J Econ* 106(2):407–443
- Becker G (1964) Human capital. Columbia University Press for the National Bureau of Economic Research, New York
- Bell D (1973) The coming of post-industrial society: a venture in social forecasting. Basic Books, New York
- Berry CR, Glaeser EL (2005) The divergence of human capital levels across cities. NBER working paper no. 11617 Sep 2005
- Caves RE (2000) Creative industries. Harvard University Press, Cambridge, MA
- Clark TN, Lloyd R, Wong KK, Jain P (2002) Amenities drive urban growth. *J Urban Aff* 24(5):493–515
- Drucker P (1993) Post-capitalist society. HarperCollins, New York
- Feser EJ (2003) What regions do rather than make: a proposed set of knowledge-based occupation clusters. *Urban Stud* 40(10):1937–1958
- Fischer MM (2011) A spatial Mankiw, Romer and Weil model: theory and evidence. *Ann Reg Sci* 47(2):419–436
- Florida R (2002) The rise of the creative class. Basic Books, New York
- Florida R (2008) Who is your city? Random House, New York
- Florida R (2012) The rise of the creative class: revisited. Basic Books, New York
- Gabe T (2009) Knowledge and earnings. *J Reg Sci* 49:439–457
- Glaeser EL, Kolko J, Saiz A (2001) Consumer city. *J Econ Geogr* 1:27–50
- Glaeser EL, Saiz A (2003) The rise of the skilled city. Brookings-Wharton Pap Urban Aff 5:47–94
- Glaeser EL (2011) Triumph of the city: how our greatest invention makes us richer, smarter, greener, healthier and happier. Penguin Press, New York
- Inglehart R (1989) Culture shifts in advanced industrial society. Princeton University Press, Princeton
- Jacobs J (1969) The economies of cities. Random House, New York
- Lucas R (1988) On the mechanics of economic development. *J Monet Econ* 22:3–42
- Machlup F (1962) The production and distribution of knowledge in the United States. Princeton University Press, Princeton
- Mankiw NG, Romer D, Weil D (1992) A contribution to the empirics of economic growth. *Q J Econ* 152:407–437
- Massey D (1984) Spatial divisions of labor: social structures and the geography of production. Methuen, New York
- McCann P (2001) Urban and regional economics. Oxford University Press, Oxford
- McGranahan D, Wojan T (2007) Recasting the creative class to examine growth processes in rural and urban counties. *Reg Stud* 41(2):197–216
- Mincer J (1974) Schooling, experience and earnings. Columbia University Press for the National Bureau of Economic Research, New York
- Piore MJ, Sabel CF (1984) The second industrial divide: possibilities for prosperity. Basic Books, New York
- Porter ME (1990) The competitive advantage of nations. Free Press, New York
- Rauch J (1993) Productivity gains from geographic concentration of human capital: evidence from the cities. *J Urban Econ* 34:380–400

- Roback J (1982) Wages, rents, and the quality of life. *J Political Econ* 90(6):1257–1278
- Saxenian A (1994) Regional advantage: culture and competition in Silicon Valley and route 128. Harvard University Press, Cambridge, MA
- Scott AJ (2009) Human capital resources and requirements across the metropolitan hierarchy of the USA. *J Econ Geogr* 9:207–226
- Smith A (1776) The wealth of nations. Random House, New York (2000)
- Sternberg RJ, Lubart TI (1999) The concept of creativity: prospects and paradigms. In: Sternberg RJ (ed) *Handbook of creativity*. Cambridge University Press, New York
- Ullman EL (1958) Regional development and the geography of concentration. *Pap Proc Reg Sci Assoc* 4:179–198
- Vernon R (1966) International investment and international trade in the product cycle. *Q J Econ* 2:190–207

Arthur Grimes

Contents

18.1	Introduction	332
18.2	Terminology: Infrastructure, Regions, and Growth	333
18.3	Infrastructure and Growth: A Spatial Equilibrium Model	333
18.4	Causality and Spatial Spillovers	340
18.5	Infrastructure as a Real Option	342
18.6	Conclusions	349
	References	351

Abstract

This chapter outlines two models for analyzing the relationship between infrastructure and regional growth and discusses relevant empirical examples. The first model adopts a standard spatial equilibrium approach and shows that the effect of new infrastructure on regional activity depends on its direct impacts on local productivity, local amenities, and the price of non-traded goods, especially housing. These impacts are determined, in part, by how existing characteristics of the region complement the specific investment. If infrastructure contributes positively to real amenity-adjusted net wages, the local region increases its attractiveness, and the result is an influx of firms and individuals to the region. In turn, this has dynamic effects that may amplify or attenuate the initial growth impetus. It is also possible that an infrastructure project contributes negatively to real amenity-adjusted net wages, imparting a negative influence on equilibrium regional activity. The second model treats

A. Grimes

Motu Economic and Public Policy Research, Wellington, New Zealand

University of Auckland, Auckland, New Zealand

e-mail: arthur.grimes@motu.org.nz

a major infrastructure investment as a real option that gives private sector developers the option, but not the obligation, for further development. The value of this option must be included by authorities when assessing benefits of a new infrastructure project. They need to judge the direct private sector responses to an investment plus the indirect equilibrium responses under alternative states of nature. The model shows that for a major infrastructure project, as in the case of other real options, a certainty equivalent approach is generally inadequate for investment analysis since that approach may underestimate the benefit of a new project when future states are uncertain, learning occurs, and decision-making is sequential.

18.1 Introduction

Investment in major infrastructure items is costly. Costs are predominantly borne up front, while benefits may not be apparent for some years after the project's commencement. Benefits may, however, last for many decades. They can accrue to the region's initial inhabitants and may induce migration flows so accruing also to people shifting from outside the region. Moreover, there may be spatial spillovers in the form of positive or negative impacts on other regions, particularly nearby ones.

The scale, and even the nature, of benefits is frequently difficult to assess *ex ante*, so there are risks both of underinvestment (especially if investment authorities are risk averse) and of overinvestment ("white elephant" projects). Given the difficulties in planning major infrastructure investments, international comparisons (such as the World Economic Forum's *Global Competitiveness Report*) show that regions and countries have widely disparate standards of infrastructure.

This chapter sets out two analytical frameworks that can be used to assess impacts of infrastructure investments on regional growth outcomes. Examples of empirical literature that illustrate aspects of the frameworks are included. The two frameworks correspond to differing treatments of uncertainty. The first is a framework based on spatial equilibrium. It details key responses that must be assessed when analyzing the impact on regional growth of a new infrastructure investment. These include direct and indirect impacts of the new infrastructure on the level and growth of regional activity and population.

The spatial equilibrium approach is essentially a certainty equivalent approach with clearly defined responses of variables to the investment. The second framework, based on real options theory, is relevant in a world with uncertainty, learning, and sequential decision-making. In these circumstances, an infrastructure investment can provide an option, but not an obligation, for future development to contribute to regional growth. An illustrative model is outlined showing the implications of these conditions for infrastructure investment decisions and how these decisions may differ from those derived from a certainty equivalent cost benefit analysis.

We define key terms in the next section prior to describing the two frameworks. A short set of observations linking the two frameworks concludes.

18.2 Terminology: Infrastructure, Regions, and Growth

Infrastructure is a widely used, but imprecise, term. Gramlich (1994) defines infrastructure as “large capital intensive natural monopolies such as highways, other transportation facilities, water and sewer lines, and communications systems.” Most examples are publicly owned, but some are owned privately. Broader definitions of infrastructure exist that include human capital investment and research and development (R&D) capital. The term is used in this chapter to include capital-intensive investments that service multiple users. The definition is extended beyond Gramlich’s narrow definition to include large-scale capital-intensive amenity investments (e.g., a major waterfront redevelopment or a hospital) and large-scale knowledge centers such as a major university.

Region describes a spatial unit that may or may not correspond to an administrative unit. A region may refer to a country within a larger international grouping (e.g., Denmark within Europe). A state or administratively defined city may constitute a region within a country. An urban area or labor market area (“labor shed”) may constitute a region that crosses administrative boundaries (and the region may even stretch across a country border). The focus here is on economic rather than administrative relationships. Impacts of infrastructure investment are more likely to spill over across administrative rather than economic boundaries (e.g., a large county-funded parking building may affect commuters from surrounding counties).

In analyzing *growth*, we must define the measure that is growing and define the time horizon over which growth proceeds. Regional growth analysis typically deals either with production or employment (with the latter alternatively measured by labor force or population) as the measure that is subject to growth. Generally, all of these measures will be growing jointly. The time horizon is important for distinguishing between a level shift in activity (achieved over some finite-time period) and a step shift in the long-term rate of economic growth associated with endogenous growth mechanisms (Aghion and Howitt 1998). It is frequently difficult to distinguish empirically between the two cases. This is particularly so where amplification mechanisms exist in the case of a level shift that, over a long period, cause a multiplicative shift in activity relative to the initial shock, but nevertheless leave the long-term growth rate unchanged once a new higher equilibrium is attained.

18.3 Infrastructure and Growth: A Spatial Equilibrium Model

The concept of spatial equilibrium is at the heart of regional science analysis. Roback (1982), building on Rosen (1979), provided a model in which agents (firms and individuals) choose between locations to maximize their profit or utility. Firms trade off productivity against labor and rental costs to maximize profits; individuals trade off amenities and income (wages) against living costs, especially

costs of non-tradables such as housing. The resulting migration of firms and individuals across regions leads to changes in wage and land costs that, in equilibrium, equate the net benefits of locating in one region to those in each other region. A change in any variable that affects productivity and/or amenities, such as a new infrastructure investment, results in a change in population and prices so as to restore spatial equilibrium.

The spatial equilibrium concept can be summarized, as in Glaeser and Gottlieb (2009), by assuming that the utility level of individuals in any region i , $U(G_T^i, G_N^i, \theta^i)$, is determined by their consumption of traded goods (G_T^i), non-traded goods (G_N^i), and local amenities in region i (θ^i). Given the individual's budget constraint and the assumptions that traded goods prices follow the law of one price and that non-earned income is unaffected by location, utility can be expressed by the indirect utility function, $V(Y^i, P^i, \theta^i)$, where Y^i is locally earned wage income and P^i is the price of non-traded goods (including housing services). Both Y^i and P^i are endogenous and so reflect the population and productive characteristics of a region. In turn, these features will be influenced by the level and type of infrastructure within a region. In a spatial equilibrium, $V(Y^i, P^i, \theta^i) = \bar{U}$, where \bar{U} is the (equal) level of utility that would be obtained by locating in any other region.

We adopt a generic spatial equilibrium model to link regional growth outcomes to infrastructure. The generic model builds on the reduced form model of Overman et al. (2010), with the role of infrastructure made explicit. We concentrate on outcomes for a specific region, holding constant outcomes for other regions. The model can be extended to multiple regions to take account of the possibility that other regions' outcomes may be affected by decisions in the region under study. If the region under consideration is small relative to the others, these spillover effects may be small. Hence, for expositional purposes, the spillover complication to the simple model is unnecessary; we return to this issue in interpreting the model's properties and predictions.

The model comprises three reduced form relationships – determining wages, house prices, and amenity benefits, respectively. Let w be the after-tax wage received by an individual in the region, L be the region's labor force (which is assumed proportional to the region's population), and I be the level of infrastructure available in the region. In interpreting the model, we will at times treat each of L and I as a vector of alternative types of labor and infrastructure inputs, respectively, but the basic model treats each as a scalar.

The first relationship, reflecting the regional production function, is given by

$$w = W(L, I) \quad (18.1)$$

The signs of the partial derivatives (W_L and W_I) in Eq. (18.1) are both indeterminate. (Unless otherwise stated, partial derivatives are evaluated within the neighborhood of the initial equilibrium position.) A diminishing returns production environment has $W_L < 0$ since full employment can, ceteris paribus, only be achieved with lower wages given a larger labor force. Conversely, $W_L > 0$ for

a region with increasing returns to scale. This may be the case, for example, within a city that has unexploited agglomeration externalities. The sign of W_I is determined by the productive benefits of a new infrastructure project relative to the cost of the project. A new project that has costs higher than its benefits (i.e., a benefit cost ratio of less than one) and that is financed from within the region (either through added taxes or by raising debt) will cause w to fall (Coleman and Grimes 2010), that is, $W_I < 0$. An investment with a benefit cost ratio in excess of one has $W_I > 0$.

The second relationship relates house prices (or, more generally, prices of non-traded goods and services, which also includes commuting costs), denoted h to the labor force and to the level of infrastructure:

$$h = H(L, I) \quad (18.2)$$

We assume that $H_L > 0$ so that greater population places pressure on land and house prices for a given quality, where location of a property in relation to the CBD and amenities is included as a quality attribute. Greater population causes a city to expand if there is a new supply of lower priced housing on the periphery, but the true housing cost (embodying both direct housing costs plus transport costs to work and amenities) is assumed to rise even where the direct (quality-unadjusted) housing price falls. By contrast, we assume that $H_I < 0$, so that a new infrastructure project reduces the true housing cost at a given location once changes in accessibility costs are included. An example is a new transport route that reduces the cost of commuting from an existing location to work or that improves access to amenities. Another example is a new broadband connection that reduces the cost of accessing some services.

The third relationship relates amenity benefits, a , to labor force and infrastructure provision:

$$a = A(L, I) \quad (18.3)$$

The sign of A_L is indeterminate. A larger population facilitates improved amenities through increasing returns to scale. For example, a larger city may be able to afford an opera house or professional sports team, whereas a smaller city is unable to do so. However, a larger population may result in congestion externalities that reduce the amenity value of a city. There is also a possibility that a larger and more diverse population increases innovation and technological change and thereby the long-run growth rate of the region. Many new infrastructure investments will enhance amenity benefits for individuals (e.g., a better road to the beach, a new opera house, or improved broadband connectivity); for these investments, $A_I > 0$. However, some forms of new infrastructure investment (such as a new airport runway) will reduce amenity benefits for local residents and, for these investments, $A_I < 0$. For simplicity, the discussion below assumes that $A_I > 0$, but the model can equally be used to analyze the impacts of an investment that reduces amenity values.

Following Overman et al. (2010), an individual's utility, u , is defined as "amenity-adjusted net wages" deflated by living costs:

$$u = \frac{aw}{h} = \frac{A(L, I)W(L, I)}{H(L, I)} \quad (18.4)$$

In spatial equilibrium, $u = \bar{u}$, where \bar{u} is the utility that can be gained in a reference region. We hold \bar{u} constant in the analysis that follows, although this assumption can be relaxed in both theoretical and empirical work. Given the requirement that $u = \bar{u}$ in equilibrium, any infrastructure investment that alters any of a , w , or h must have an offsetting equilibrium effect on one or more of the other two variables. For instance, an infrastructure project that directly raises productivity, and hence w , must result in either a reduction in amenity values, a , or an increase in local living costs, h , at the new equilibrium.

To demonstrate this formally, differentiate u with respect to I and set the result equal to zero (since \bar{u} remains unchanged following a change in the region's level of infrastructure). In doing so, we treat L as endogenous. Given that we are interested in the growth effects of a change in I , our focus is on dL/dI , the change in labor force for a small change in infrastructure. The result is shown in Eq. (18.5):

$$\frac{dL}{dI} = \frac{\frac{w}{h}A_I + \frac{a}{h}W_I - \frac{aw}{h^2}H_I}{\frac{aw}{h^2}H_L - \frac{w}{h}A_L - \frac{a}{h}W_L} \quad (18.5)$$

Expression (18.5) enables us to analyze a variety of channels through which infrastructure affects regional activity. Assume initially that diminishing returns to scale exist in production and that extra population causes net congestion costs; thus, each of W_L and A_L is negative. In that case, the denominator of Eq. (18.5) is positive. Furthermore, if the productive benefits of the new infrastructure outweigh its costs, so that W_I is positive, then the numerator is also unambiguously positive. Under these conditions, an increase in infrastructure results in an increase to the local labor force (and hence population and activity) in the region.

We can relax the assumption that W_I is positive and still obtain the same (signed) result provided that A_I and/or H_I are sufficiently large in absolute value to leave the numerator positive. Thus, even if the infrastructure is not net positive for local production, provided it increases amenity values or decreases living costs sufficiently, it will still raise utility and so induce a population inflow to the area. Examples of such infrastructure projects may be a waterfront redevelopment (improved amenity) or a new transport link between housing and work locations (i.e., a reduction in the full cost of housing). The ensuing improvement in amenities, or reduction in housing costs, leads to a larger labor force through net inward migration; the required increased production to employ the larger workforce is achieved through a lower wage rate. In the case of improved amenities, people accept the lower wages so that they can remain living in a more desirable region. In the case of reduced housing costs following an infrastructure improvement, the

reduction in housing costs enables wages to fall while leaving people with the same real incomes as before the investment.

The corollary of this analysis is that a poorly chosen infrastructure project (i.e., one that has costs exceeding its productive, amenity plus housing benefits) will result in a *decline* in local population despite its potential to service a larger population. This result occurs since the cost of servicing the project exceeds the benefits it brings. This inefficiency case is particularly germane where all costs are borne by the local region and so are reflected in disposable incomes, w . As an example, Siegfried and Zimbalist (2000) provide evidence that the construction of large sports stadia in the United States may have caused declines in local activity and local real incomes after controlling for other factors affecting regional outcomes. This example also illustrates the potential for an infrastructure investment to affect the composition of population in a region through a Tiebout-sorting mechanism (e.g., sports fans may relocate to be near a stadium, whereas other citizens may depart the local area). While potentially important, we abstract from sorting issues in our analysis.

If, for an inefficient project, the infrastructure costs are borne elsewhere (e.g., by central government), the benefits are internalized within the region, but the costs are not internalized to that region and so are not reflected through a reduction in w . The result in this case is an expansion in L , but this expansion is inefficient due to the lack of internalization of costs to the region. Belich (2009) documents substantive cases of inefficient infrastructure investments (canals and railroads) in the new settler colonies, such as Canada, in the nineteenth century, with adverse economic and financial effects.

We now turn attention to the denominator of Eq. (18.5), assuming that the numerator is positive. Differing parameters in the denominator have the potential to cause significant multiplicative dynamic effects following an infrastructure investment. Consider an economy that exhibits increasing returns to scale relative to one with constant returns. In the increasing returns case, W_L is positive; this may be due to a larger labor force lifting overall wages through agglomeration effects (Fujita et al. 1999). Provided W_L is not “too large,” the denominator remains positive but declines in size relative to the constant returns to scale case, so that dL/dI increases. Thus, regional growth in response to an infrastructure investment is greater where agglomeration forces (positive W_L) are at work compared to an economy with no such forces. The importance of this aspect has been emphasized, especially in the United Kingdom, in recent official approaches to transport project assessment techniques (e.g., Eddington 2006).

A similar result occurs if net amenity benefits are positively linked to region size (i.e., positive A_L). In this case, the improvement in the range or quality of amenities as a region grows outweighs congestion costs of that growth. An increase in infrastructure provision supports a larger population, and this, in turn, enhances the attractiveness of the region which spurs further population growth.

A highly elastic housing supply (low absolute H_L) facilitates a larger population response to an infrastructure investment compared with the case where housing supply is limited and prices rise sharply as population expands (Glaeser et al. 2005).

Thus, there is a complementarity between new infrastructure provision and flexible zoning and construction regulations that enable a region to grow strongly in response to new infrastructure.

From Eq. (18.5), higher values of A_L and W_L , and/or a lower value of H_L , induce a heightened increase in u , leading to a greater inward migration response. Provided the denominator of Eq. (18.5) is positive, the offsetting contributions of the three factors that influence amenity-adjusted real wages ensure that the change in population following an infrastructure project is bounded. For instance, an increase in utility arising from improved amenities following a new infrastructure investment will be offset by an increase in housing costs and/or a decrease in wages as population and activity increases.

Population adjustment following an infrastructure project will, in practice, not be completed immediately. Instead, we may postulate a migration function such as

$$m = M(u - \bar{u}) \quad (18.6)$$

where m is defined as dL/dt , t represents time, and $M_u > 0$. Thus, people and firms migrate to a region for as long as location in that region yields higher utility than locating elsewhere. The partial adjustment mechanism in Eq. (18.6) may reflect nonlinear costs of adjustment. Expression (18.6) translates the equilibrium adjustment response analyzed above into a regional growth response. For a given investment project and a given migration sensitivity, the larger is dL/dI , the larger and the longer will be the growth phase following the project, until a new equilibrium is reached. Once this new equilibrium is reached, the growth rate will return to the initial path that existed prior to the infrastructure investment, albeit with activity and population at higher levels.

The analysis has so far considered only the stable case, where the denominator of Eq. (18.5) is positive, so that there are offsetting gains and losses from an expansion that cause the regional response to an infrastructure project to be bounded in the neighborhood of the initial equilibrium. If, instead, the denominator is negative, there are insufficient offsetting forces to limit region size within the neighborhood of the initial equilibrium. Thus, utility keeps rising as population expands in response to an initial upward shift in infrastructure provision.

This situation corresponds to one of explosive growth which continues until the population has grown sufficiently to bring about offsetting forces to curb the growth. Thus, the initial response to a new project may be a population influx that induces amenity and productive benefits that are not outweighed by steeply rising housing costs. This may be because the infrastructure has been planned so as to service a large area of developable land. However, once this developable land is fully utilized and/or once congestion costs rise as the infrastructure is used intensively, then H_L rises and/or A_L falls, so that dL/dI is once again stable and u is eventually equated to \bar{u} .

During the intervening period, the migration response in Eq. (18.6) operates, and a prolonged regional growth period may arise following an initial investment. The long-term equilibrium result is a “spikey” landscape in terms of concentrations of

infrastructure and economic activity (Florida et al. 2008). Some tightly defined regions grow to have dense concentrations of population and economic activity serviced by a comprehensive set of infrastructures, while others have low population density, sparse economic activity, and little infrastructure.

Shenzhen, in the Pearl River delta of Southeastern China, is a recent example of explosive development coupled with substantive provision of publicly provided infrastructure, followed by more gradual growth. The city expanded from a population of 350,000 in 1982 and 1,200,000 in 1990 to 7,000,000 in 2000. Its growth then slowed to reach approximately 10,000,000 in 2010. While still growing in the decade to 2010, the absolute and percentage growth rates both fell relative to the preceding decade reflecting the ending, or at least the slowing, of the explosive growth period.

The type of infrastructure, as well as its overall level, can have an important bearing on regional growth patterns. It is helpful to consider I in our model as a vector of infrastructure inputs, each element of which has differing direct (and indirect) impacts on wages, amenities, and house prices. An infrastructure investment designed to service a new industrial mill may be consistent with a diminishing returns to scale environment ($W_L < 0$). By contrast, the presence of a research university is positively associated with high value R&D and patenting activity locally (O'hUallachain 1999). This is particularly the case in a city with strong availability of skilled labor (Ceh 2001). High R&D activity and a high proportion of skilled labor, in turn, may be associated with strong regional growth through a variety of endogenous growth mechanisms (Aghion and Howitt 1998) so that $W_L > 0$.

The importance for regional growth outcomes of complementarities between the new infrastructure and the factors already in the region is highlighted also by Forman et al. (2012). Advanced Internet provision in the USA raised incomes in counties that already had high incomes, an educated workforce, dense populations, and IT-intensive industries. It had no effect on incomes in counties without those characteristics. Thus, one cannot simply extrapolate the findings of an infrastructure's impact on one set of regional outcomes to other regions. The characteristics of the region influence the sign (and size) of the key parameters in Eq. (18.5), and the impacts even of the same type of infrastructure can differ markedly across regions according to regional characteristics.

One of the key insights of the spatial equilibrium model is that any increase in amenity-adjusted wages following an infrastructure project must be matched by a proportionate increase in local prices, including land. Haughwout (2002) demonstrates that, under certain conditions, the aggregate increase in regional land values attributable to an infrastructure project is the appropriate measure of the benefits of that project. The fixed factor (land) reflects all net benefits of the project since the net return to each mobile factor (labor and capital), and the price of traded goods are all set externally and so do not change in response to the local infrastructure investment.

This approach to measuring benefits of an infrastructure project captures amenity as well as productive benefits. It therefore provides a more complete approach

to measuring benefits than captured by earlier methods that relied on regional aggregate production functions (Aschauer 1989) or regional aggregate cost functions (Morrison and Schwartz 1996). These earlier approaches are unable to capture the benefits of new infrastructure that accrue through all amenity, production (wage), and housing channels. Similar difficulties are faced in conventional cost benefit analyses that sum postulated benefits to arrive at an estimated benefit total. An example of the land value approach to measuring benefits of a new infrastructure investment is Grimes and Liang (2010) who analyzed the net value attributable to a new motorway link. In their application, the land value approach gives a considerably higher estimate of value creation than does a conventional transport cost benefit analysis.

18.4 Causality and Spatial Spillovers

The generic model developed above is based on spatial equilibrium relationships. A key issue in the analysis of the linkage between infrastructure and regional growth is whether the relationship between these two variables is purely associative or whether it is causal and, if it is causal, which is the direction of causality.

To understand the importance of this issue, consider a set of regions that are in spatial equilibrium, each with infrastructure at optimal levels given the features of each region (i.e., given the production relationships, level of natural amenities, and land availability). Now consider the impact of an increase in infrastructure in one region, j . If infrastructure was previously at its optimal level, an addition to region j 's infrastructure will raise costs (taxes) by more than it contributes to productivity plus amenity value. The numerator of Eq. (18.5) is negative for any level of infrastructure beyond the initial optimal level, so an infrastructure increase causes utility to fall from $u_j = \bar{u}$, leaving $u_j < \bar{u}$. The result, from Eq. (18.6), is an outflow of migrants from region j , that is, negative regional growth driven by the infrastructure investment.

Now consider a region, k , in which, initially, $u_k = \bar{u}$, but where infrastructure is at a suboptimal, constrained level. The suboptimal infrastructure level may have arisen because of a lagged response to some other factor that has raised the productive or amenity potential of the region. In this case, a new infrastructure investment is an endogenous response to the external factor that has caused the productive or amenity potential of the region to increase. The numerator of Eq. (18.5) is positive at the initial infrastructure level and so an expansion of infrastructure initially causes $u_k > \bar{u}$. In turn, this causes inward migration so that population and activity both rise in the region.

The two cases will initially be observationally equivalent in terms of migration patterns and growth (since both are in equilibrium with $u_j = u_k = \bar{u}$). However, in the former case, a new infrastructure investment causes population and activity to fall; in the latter case, a new infrastructure investment facilitates an increase in population and activity. These examples make it clear that there is no consistent causal relationship between new infrastructure and regional growth. The new

infrastructure must raise amenity-adjusted (net) wages and/or reduce housing (or other non-traded) costs to induce regional growth. Even in the second case (region k), it is questionable whether the new infrastructure can be said to *cause* regional growth. Regional growth would certainly be constrained without the new infrastructure, but it is best to consider the increase in infrastructure as codetermined with regional activity, rather than causing activity to rise.

The endogeneity of infrastructure makes it difficult to extract empirical estimates of the impact of an infrastructure investment on regional growth outcomes. Wu and Gopinath (2008) attempt to do so by instrumenting the endogenous variables (including infrastructure) in their three-stage least squares spatial model. In accordance with the previous analysis, they show that treating road density (their infrastructure measure) either as exogenous or endogenous has material effects on estimates of the impacts of other factors on regional wage, price, and activity outcomes.

The key practical difficulty in using an instrumental variables approach is to find strong instruments that affect not only the level of regional infrastructure but also its change over time. In this respect, models relying solely on fixed regional characteristics (such as climate or lagged road density) that may predict the *level* of infrastructure across regions do not generally contain sufficient information to predict *changes* in infrastructure. From a policy perspective, however, it is the impact of a change in infrastructure provision that is of interest. Use of exogenous instruments that contain time variation (e.g., Wu and Gopinath's use of an "agricultural net return" variable) provide greater promise as instruments in this respect than fixed regional characteristics.

The Wu and Gopinath analysis is useful also in that they test for spatial lags in the relationship between regional outcomes and infrastructure investments. They find positive spatial autocorrelation in human capital supply, developed area supply and demand, and road density. Thus, a major infrastructure development in one region spills over positively into neighboring regions for these variables. It may nevertheless be the case elsewhere that spatial lags are negative for certain types of investment (i.e., a displacement effect of infrastructure on activity). Hence, we cannot generalize about the spatial impact of infrastructure investment on neighboring regions' growth from analysis of particular examples.

An alternative approach to analyzing the causal impact of infrastructure investment is to examine the impacts of a natural experiment involving an infrastructure project. One such situation is where new infrastructure has been constructed for a purpose that is exogenous to the region concerned. Coleman (2012) examines the impacts of the Erie Canal on rural New York population and activity. The canal was constructed primarily to link New York City with Chicago and the Great Lakes, not specifically to service rural New York communities, so is a natural experiment with respect to these latter communities. The analysis shows that the decreased cost of transport due to the canal caused (a) existing households to switch gradually from home production, (b) an influx of migrants who were more oriented toward market production than incumbent households, and (c) a delayed copying by incumbents of migrants' production patterns. In this case, the relevant characteristics of the vector

L are the sources of the population, with knowledge and/or skills varying according to prior location. The influx of migrants and changes in market production following the canal's completion are consistent with the spatial equilibrium approach coupled with a partial adjustment mechanism. Given the exogeneity of the canal to these rural communities, the canal can be regarded as having a causal role in determining the patterns of growth in these regions.

Another natural experiment is documented by Duflo and Pande (2007) who examine productivity impacts of large irrigation dams in India. The analysis uses river gradient as an instrument for the choice of dam site, and contrasts outcomes above and below a dam following its construction. Downstream agricultural production increases, and its volatility falls, as a result of a dam. In contrast, upstream production shows no significant increase, whereas its volatility rises, causing increased hardship for upstream farmers. Given the exogenous choice of dam location, both impacts can be considered causal responses of production to the dam's construction. The negative impact of a dam on upstream farmers is an example of a negative spatial spillover that needs to be accounted for in assessing the overall benefits of the infrastructure project.

18.5 Infrastructure as a Real Option

The analysis in previous sections is formulated in terms of a deterministic or certainty equivalent concept of equilibrium. As an *ex post* concept, this approach is useful for understanding how variables have adjusted to infrastructure investments. The approach may also be appropriate for *ex ante* consideration of the regional growth impacts of small infrastructure projects with well-defined outcomes. However, most major infrastructure projects are built to last a long time and the future environment as well as future outcomes may be highly uncertain. *Ex ante*, therefore, it is not just the size but also the nature of benefits that may be difficult to determine.

For instance, public authorities may have reasonable *ex ante* estimates of the productivity impacts of a small infrastructure investment that is closely coordinated with a specific private sector development. In this case, the authority's estimates of responses, such as W_l in Eq. (18.5), may be sufficient to underpin its investment decision. However, the quality of information about potential private sector responses may be much less complete for infrastructure investments that are capable of inducing broader impacts. An example is the difficulty in predicting firm and household responses to widespread public provision of fiber-optic cable for ultrafast broadband services (Grimes et al. 2012). The Erie Canal project, cited in Sect. 18.4, is an example in which the canal-building authorities were unlikely to have known in advance that production would change in rural New York communities as a result of building the canal.

In such circumstances, a major infrastructure investment may create a real option for alternative development paths (Miller and Lessard 2008). As is the case for a financial option, there is no obligation for (private or public) developers

to make any specific use of the development option, but the value of the option may nevertheless materially influence the decision of whether or not to invest.

In analyzing infrastructure projects that create new development opportunities, we apply techniques derived from real options analysis (Dixit and Pindyck 1994; Guthrie 2009). Dixit and Pindyck show that a present discounted value analysis of cash flows (or of utilities within a broader cost benefit framework, as adopted in Sect. 18.3) may seriously distort investment decisions relative to the optimum under certain circumstances. The key combination of circumstances that makes real options analysis indispensable is where (a) uncertainty exists about future states of the world, (b) learning occurs about these states over time, and (c) future decisions can be made based on updated information about these states. The example that follows illustrates insights that can be gained through real options analysis. While simple, it demonstrates how incorporation of real options analysis can substantively change infrastructure investment decisions under conditions of uncertainty relative to decisions made under the certainty equivalent framework adopted in Sect. 18.3. As in Sect. 18.3, we assume that the public infrastructure investor adopts the perspective of social planner and wishes to invest where aggregate benefits exceed aggregate costs so that utilities are raised.

Consider a public sector infrastructure provider that has the opportunity to build a bridge that would open up a large new area for development. We adopt a three-period model ($t = 0, 1, 2$) in which the bridge can be built in $t = 0$, or not at all (“period” may refer to multiple years). This timing restriction is made so that we do not have to consider the potential option value of delaying the infrastructure investment (Dixit and Pindyck 1994); nevertheless, the generality of the example transfers to a model where timing of the infrastructure project is also a choice variable.

In each period, there is uncertainty about future states of the world, but learning occurs as time proceeds. If the bridge is not built, the (utility) payoff in each of $t = 0, 1, 2$ is normalized to 0. The cost of the bridge is X and, assuming no immediate benefits during construction of the bridge, the net payoff in $t = 0$ is $-X$ if the bridge is built. Henceforth, we consider payoffs on the assumption that the bridge is built. We assume that the private sector can develop the new area at $t = 1$ if and only if the bridge is built.

If the private sector proceeds with development in the new area following construction of the bridge, this development will induce growth in the population, labor force, and employment of the city as analyzed in Sect. 18.3. We assume that this growth is sourced from outside of the region of analysis. Private development only occurs in $t = 1$ if it is expected to be profitable over the life of the development (i.e., over $t = 1, 2$).

We can model the resulting choices and outcomes equivalently through a (visual) decision tree approach or by adopting a payoff matrix approach. While slightly more complicated, the latter approach enables extensions to more complex analytical problems and so is the approach adopted here. Following Guthrie (2009), news about the state of the world in $t = 1$ is assumed to be either “good” or “bad.” Similarly, a second set of news emerges about the state of the world at $t = 2$ that can be either good or bad. News about $t = 1$ ($t = 2$) is only revealed during $t = 0$ ($t = 1$),

Table 18.1 Payoff matrices for multistage investment

		A: Payoffs $Y(i,t)$ for $[B,N]$		
		t (period)		
i		0	1	2
0		$Y(0,0)$	$Y(0,1)_N$	$Y(0,2)_N$
1			$Y(1,1)_N$	$Y(1,2)_N$
2				$Y(2,2)_N$

i is the number of “bad” events occurring prior to period t
 $[B,N]$ indicates that the bridge is built but no private development takes place

		B: Payoffs $Y(i,t)$ for $[B,G]$		
		t (period)		
i		0	1	2
0		$Y(0,0)$	$Y(0,1)_G$	$Y(0,2)_G$
1			$Y(1,1)_G$	$Y(1,2)_G$
2				$Y(2,2)_G$

i is the number of “bad” events occurring prior to period t
 $[B,G]$ indicates that the bridge is built and private development takes place

and so is not known at the time the bridge investment decision has to be made. The private sector development choice occurs at the start of $t = 1$ following the first set of news but prior to the second. Table 18.1 shows two payoff matrices. Table 18.1A shows payoffs by period and state of the world if the private sector does not develop. Table 18.1B shows payoffs if the private sector chooses to develop. $[B,N]$ indicates that a public sector decision to build the bridge was taken $[B]$ but that the private sector decided not to further develop the area $[N]$. $[B,G]$ indicates that the public decision to build the bridge and the private decision to develop (grow) the area $[G]$ are both taken.

Payoffs are state-dependent. The state of the world in t is summarized by i , the number of bad news events that have occurred prior to t . At $t = 0$, $i = 0$ since there have been no prior bad news events. At $t = 1$, i can take the value of zero (no bad news event during $t = 0$) or one (one bad news event). At $t = 2$, i can be zero, one, or two. The payoffs to the bridge investment in each period decline as the state of the world worsens (i.e., as the number of prior bad news events increases).

The payoff matrices have elements $Y(i,t)$, each representing the payoff corresponding to the number of prior bad news events (i) at period t . For each of the $[B,N]$ and $[B,G]$ cases, the initial period payoff is $Y(0,0) = -X$, being the cost of the bridge project. For the $[B,N]$ case (the bridge is built but no subsequent private sector development takes place), $Y(0,1)_N > Y(1,1)_N$. If private sector development occurs in $t = 1$, that is, $[B,G]$, the net payoffs (after deducting the costs of private development) in $t = 1$ have the property that $Y(0,1)_G > Y(1,1)_G$.

After $t = 1$ (i.e., after the private sector development decision has occurred), further news is revealed. Thus, in $t = 2$, there are three potential states of the world ($i = 0,1,2$). If no private sector development has occurred, the payoffs in $t = 2$ corresponding to $i = 0,1,2$ are $Y(0,2)_N$, $Y(1,2)_N$, and $Y(2,2)_N$, respectively,

where $Y(0,2)_N > Y(1,2)_N > Y(2,2)_N$. If private sector development occurs, the respective payoffs in $t = 2$ are $Y(0,2)_G$, $Y(1,2)_G$, and $Y(2,2)_G$, where $Y(0,2)_G > Y(1,2)_G > Y(2,2)_G$.

We assume that $Y(0,2)_G > Y(0,2)_N$ so that private sector development is worthwhile where two good states of nature have emerged. Furthermore, we assume that $Y(1,2)_N > Y(1,2)_G$ and $Y(2,2)_N > Y(2,2)_G$ so that private sector development is not worthwhile where at least one bad state of nature occurs. The more pronounced negative outcomes given poor states of the world when private sector development proceeds (relative to a situation where development does not proceed) reflect the extra costs involved in expanding the city boundaries through the infrastructure investment when it would have been preferable, ex post, to have remained within the original boundaries. The larger benefit attributable to development when two good states have occurred ($i = 0$) reflects positive returns to expansion when that expansion is supported by a sequence of good news events.

A brief note is useful here about the specification of the payoffs. Each payoff is deterministic once the (i,t) combination is revealed. For instance, the payoff for $[B,N]$ with two bad outcomes is known to be $Y(2,2)_N$. This payoff is the utility payoff (corresponding to u in the previous analysis, so accounting for amenity, productivity, and local cost impacts) at $t = 2$ in state of the world $i = 2$. The consumption capital asset pricing model (Breeden 1979) can be used to assess the utility payoff for any dollar return. A corollary of measuring the payoffs in utility terms, where each payoff is deterministic once the (i,t) combination is known, is that the real risk free rate, r , is the appropriate discount rate to use to discount future payoffs to the present (Guthrie 2009).

The payoff matrices from Table 18.1 can be used to perform a present discounted value of utility (ΣDVU) analysis, akin to a discounted cash flow analysis, as would typically be done for a cost benefit analysis. In order to do so, we need to specify the per period discount rate (r), and we must specify the probability (p) of each news event (i.e., each state transition) being bad. Table 18.2 provides an example of payoffs corresponding to the elements of Table 18.1. The net discounted utility to $t = 0$ corresponding to this example is shown beneath the table. The calculation uses a per period discount rate of $r = 0.04$ and a probability of a bad news event each period of $p = 0.5$.

In a certainty equivalent world, the analysis proceeds by calculating whether either of the $\Sigma DVUs$ is positive and, if so, choosing the path which yields the larger discounted utility. If no path yields a positive ΣDVU , then the infrastructure should not be built. In the first case (the bridge is built but no further development proceeds), the ex ante discounted utility is negative (-1.43 to two decimal places).

In the second case (both the bridge and private developments proceed), the ex ante discounted utility is again negative (-0.31). Thus, ex ante, neither a single-stage nor a two-stage development is warranted given a certainty equivalent analysis. A standard methodology will reject the bridge project because it yields a negative ΣDVU (i.e., the benefit cost ratio is less than unity), and the negative ΣDVU pertains to the infrastructure investment whether or not the subsequent private development is undertaken.

Table 18.2 Payoff matrices:
exampleA: Payoffs $Y(i,t)$ for $[B,N]$

		t (period)		
		0	1	2
i	0	-1	2	1
	1		0	-1
		2		

i is the number of “bad” events occurring prior to period t
 $[B,N]$ indicates that the bridge is built but no private development takes place

The sum of discounted utilities (ΣDVU) is given by
 $\Sigma DVU = -1.43 \cong (1)(-1) + [(0.5)(2) + (0.5)(0)]/1.04 + [(0.25)(1) + (0.5)(-1) + (0.25)(-5)]/1.04^2$

B: Payoffs $Y(i,t)$ for $[B,G]$

		t (period)		
		0	1	2
i	0	-1	1	25
	1		-1	-5
		2		
		-12		

i is the number of “bad” events occurring prior to period t
 $[B,G]$ indicates that the bridge is built and private development takes place

The sum of discounted utilities (ΣDVU) is given by
 $\Sigma DVU = -0.31 \cong (1)(-1) + [(0.5)(1) + (0.5)(-1)]/1.04 + [(0.25)(25) + (0.5)(-5) + (0.25)(-12)]/1.04^2$

However, certainty equivalent analysis misses the value of the option that the bridge investment creates. An option provides the opportunity, but not the obligation, to exercise an investment choice in future. In this example, the private developers, having observed the value of i at $t = 1$, have the option at $t = 1$ of choosing whether or not to invest in developing the area opened up by the bridge, but not the obligation to do so.

To examine the worth of the option, a value function approach is adopted, as displayed in Table 18.3, based on the values in Table 18.2. Each cell shows the forward-looking value of being in that position, given the existing value for i , at time t . In $t = 2$, the value in each cell, $V(i,t)$, is therefore identical to the equivalent payoff, $Y(i,t)$, in Table 18.2.

The value for $V(0,1)$ equals the payoff at that time, $Y(0,1)$, plus the discounted worth of the probability-weighted potential values in $t = 2$ [$V(0,2)$ and $V(1,2)$]. Thus, in the $[B,G]$ case,

$$V(0,1)_G = Y(0,1)_G + [pV(1,2)_G + (1-p)V(0,2)_G]/(1+r) \quad (18.7)$$

The same approach is adopted for each other cell at $t = 1$ in the two matrices. Similarly, $V(0,0)_G$ is calculated as the payoff at that time, $Y(0,0)$, plus the

Table 18.3 Value function matrices: example

		A: Value functions $V(i,t)$ for $[B,N]$		
		t (period)		
i		0	1	2
0		-1.43	2.00	1
1			-2.88	-1
2				-5

		B: Value functions $V(i,t)$ for $[B,G]$		
		t (period)		
i		0	1	2
0		-0.31	10.62	25
1			-9.17	-5
2				-12

i is the number of “bad” events occurring prior to period t
 $[B,N]$ indicates that the bridge is built but no private development takes place

		B: Value functions $V(i,t)$ for $[B,G]$		
		t (period)		
i		0	1	2
0		-0.31	10.62	25
1			-9.17	-5
2				-12

i is the number of “bad” events occurring prior to period t
 $[B,G]$ indicates that the bridge is built and private development takes place

discounted worth of the two probability-weighted potential value functions in $t = 1$ [$V(0,1)_G$ and $V(1,1)_G$]. Thus, in the $[B,G]$ case,

$$V(0,0)_G = Y(0,0)_G + [pY(1,1)_G + (1-p)Y(0,1)_G]/(1+r) \quad (18.8)$$

In the $[B,N]$ and $[B,G]$ cases, respectively, this approach produces a value for $V(0,0)_N = -1.43$ and $V(0,0)_G = -0.31$, each of which is identical to the sum of the discounted utilities presented below Table 18.2. Thus, the value function approach produces identical values to the ΣDVU method given the same decision-making restrictions.

However, these values are still calculated prior to taking the option into account and would only be an appropriate valuation of net benefit if the choice between private development (G) or none (N) were irrevocably committed to at the same time as the public infrastructure decision occurs ($t = 0$). In fact, the private sector choice is made only after new information has emerged at $t = 1$.

To see how this timing affects the analysis, consider the nature of the private development decision at $t = 1$ if the bridge had been built at $t = 0$. If $i = 0$, the developer has $V(0,1)_G = 10.62$ if the development proceeds, while $V(0,1)_N = 2.00$ if there is no development. Thus, if $i = 0$ at $t = 1$, the developer will maximize value by developing since $10.62 > 2.00$. Denote the resulting value (in italics) as $V(0,1)_G$.

If $i = 1$ at $t = 1$, $V(1,1)_G = -9.17$ if development proceeds, while $V(1,1)_N = -2.88$ if there is no further development. Thus, if $i = 1$ at $t = 1$, the

Table 18.4 Value function matrices incorporating option

A: Value functions $V(i,t)$: general		
i	t (period)	
	0	1
0	$V(0,0)$	$V(0,1)_G$
1		$V(1,1)_N$

i is the number of “bad” events occurring prior to period t

B: Value functions $V(i,t)$: example		
i	t (period)	
	0	1
0	2.72	10.62
1		-2.88

i is the number of “bad” events occurring prior to period t

developer will maximize value by choosing not to develop since $-2.88 > -9.17$. Denote the resulting value as $V(1,1)_N$.

Viewed from $t = 0$, the infrastructure investor will optimally recognize that the decision tree will diverge depending on the value of i at $t = 1$. If good news occurs after the bridge is built, the private development will take place, but no development will be undertaken if bad news occurs at $t = 1$. The two relevant values in $t = 1$ are therefore shown in italics in [Table 18.3](#), one in each of the $[B,N]$ and $[B,G]$ cases. [Table 18.4](#) combines these two values into a single table as viewed from $t = 0$ (dropping $t = 2$ since that period’s values are incorporated into the values at $t = 1$). [Table 18.4A](#) provides the general case, and [Table 18.4B](#) adopts the values based on [Tables 18.2](#) and [18.3](#). The actual value of the project at $t = 0$ if the bridge is built, $V(0,0)$, will then be equal to the payoff in $t = 0$, $Y(0,0)$, plus the discounted worth of the two probability-weighted value functions in $t = 1$ being $V(0,1)_G$, and $V(1,1)_N$. In the example, this yields $V(0,0) = 2.72$.

The resulting project value, $V(0,0)$, is positive, and thus the infrastructure project should proceed. The choice to build the bridge is therefore optimal (given the values in the example), but this decision is contrary to that derived from a certainty equivalent analysis both for the case where it is assumed that subsequent private development occurs and for the case where it is assumed that no subsequent private development occurs. The optimal approach incorporates the value of the real option created by the infrastructure investment, whereas the certainty equivalent approach ignores this option value. Where multistage decision-making occurs under conditions of uncertainty, coupled with the ability to update both information and development decisions after the infrastructure is built, the analytical approach must therefore extend beyond a traditional certainty equivalent analysis to one that incorporates the option benefit created by the infrastructure investment. In practice, the larger the potential breadth of impact of the new infrastructure, the greater will be the uncertainty surrounding the nature and size of benefits, and the greater will be the ability to adjust (private and public) decisions in future periods. Hence, the

options framework may, in practice, be most applicable to potentially “game-changing” investments.

The results derived from this analysis are instructive for interpreting the impacts of actual historical infrastructure investment decisions on regional growth outcomes. In some circumstances, an investment may be optimal to proceed with *ex ante* when viewed from $t = 0$ where the real option value is included in the assessment. However, *ex post*, subsequent information may lead to little or no further development. Thus, one is left with a “white elephant” project such as an unprofitable canal or railroad (Belich 2009).

A simple historical *ex post* analysis, undertaken with the benefit of hindsight, might conclude that the regional growth response was mediocre and that the investment choice was therefore poor. (And, of course, this may be the case for some projects!) However, the investment decision to build may nevertheless have been optimal *ex ante* even though the subsequent information made the investment unprofitable.

The infrastructure investment approach analyzed here is analogous to the assessment of whether or not to drill an oil exploration well. New (bad) information revealed after the well is drilled may show that it should not or cannot be developed to become a production well. In a good state of nature (oil is discovered) there is an opportunity, but not an obligation, to develop the well further; the choice of whether to exercise that option will depend on conditions at that time. As in the case of an underutilized infrastructure asset, an unsuccessful oil well does not necessarily signify that the decision to undertake the first stage investment was poor.

One implication of the inevitable *ex post* failure of some infrastructure projects – that were optimally chosen to proceed – is that a portfolio approach to infrastructure investments, as in a standard finance application, may be warranted. Each investment can be considered as part of a portfolio of infrastructure investments designed to raise utility and hence contribute to regional growth. It is the return (and risk) on the total portfolio which is of importance, not the outcome for any single investment. Thus, rather than considering whether each infrastructure project turned out to be warranted *ex post*, a more appropriate assessment might examine whether the portfolio of infrastructure investments has contributed positively to regional outcomes given that each project was assessed legitimately *ex ante* inclusive of option value.

18.6 Conclusions

Infrastructure is an integral factor supporting regional growth. The types of infrastructure considered here are capital-intensive items that service multiple users. Their impact on regional activity will differ by infrastructure type, and so one cannot postulate a generalized relationship between infrastructure investment and regional growth. The spatial equilibrium model establishes that a few key mechanisms determine the relationship between infrastructure investment and growth

outcomes. The effect of a new infrastructure project on activity depends on its direct impacts on local productivity, local amenities, and the price of non-traded goods, especially housing. These impacts will be determined, in part, by existing regional characteristics and how they complement the specific investment.

If infrastructure contributes positively to real amenity-adjusted net wages, the local region becomes more attractive and there is an influx of firms and individuals to the region. This, in turn, has dynamic effects that may amplify or attenuate the initial growth impetus. The result is, at least, a finite period of raised regional growth until a new spatial equilibrium is attained. Certain investments (particularly knowledge-based investments) that are associated with endogenous growth mechanisms may lift the regional growth rate permanently, or for a long horizon. It is possible also that an infrastructure project contributes negatively to real amenity-adjusted net wages, particularly in cases where its benefit cost ratio is less than unity and the costs are met from the local region. In this case, the investment imparts a negative impetus to regional activity.

In light of these equilibrium outcomes, it is important to analyze decision-making processes for major investments. Given that decisions on public infrastructure investments are generally made prior to the full nature of private sector development being revealed, the public authorities must assess how private investors will respond to the investment given a range of circumstances (states of nature). New infrastructure creates an option, but not an obligation, for subsequent private sector development to occur. Indeed, the identity of future investors will often be unknowable (and hence uncontractable) since some agents may not even be in existence at the time the infrastructure decision is made.

The real options approach to infrastructure investment is conceptually appropriate in these circumstances. The authorities must include an assessment of the option value that they create when making decisions on a new infrastructure project. In doing so, they must judge the private sector responses to an investment (i.e., the W_I , A_I , and H_I from the spatial equilibrium model) plus the indirect equilibrium responses (W_L , A_L , and H_L) under alternative states of nature.

A certainty equivalent approach, as in standard cost benefit analysis, is conceptually inappropriate for this analysis since that approach may underestimate the benefit of a new project when future states are uncertain, learning occurs, and decision-making is sequential. However, the real options approach has greater information demands than the certainty equivalent approach, requiring information on the range and probabilities of potential equilibrium outcomes as opposed to requiring just expected values. Given these information demands under the real options approach, scenario analysis in which alternative distributions of potential outcomes are postulated can be a useful tool to test the robustness of the infrastructure investment decision.

No matter whether the certainty equivalent or real options approach is used, *ex ante*, to evaluate a major infrastructure investment choice, decision-making for such a project is inevitably a complex task. Even the certainty equivalent approach requires policy-makers to estimate indirect as well as direct impacts of an initial infrastructure investment on a region. Both approaches

nevertheless emphasize the importance of considering regional impacts of infrastructure in a systemic manner, and emphasize also the potential importance of infrastructure investments in supporting, and providing options for, regional growth outcomes.

Acknowledgements I wish to thank Jacques Poot, Manfred Fischer, and Andrew Coleman for their helpful comments while preparing this chapter; however, all views (and any errors or omissions) are solely attributable to the author.

References

- Aghion P, Howitt P (1998) *Endogenous growth theory*. MIT Press, Cambridge, MA
- Aschauer D (1989) Is public expenditure productive? *J Monet Econ* 23(2):177–200
- Belich J (2009) *Replenishing the earth: the settler revolution and the rise of the Anglo-world, 1783–1939*. Oxford University Press, Oxford
- Breeden D (1979) An intertemporal asset pricing model with stochastic consumption and investment opportunities. *J Financ Econ* 7(3):265–296
- Ceh B (2001) Regional innovation potential in the United States: evidence of spatial transformation. *Pap Reg Sci* 80(3):297–316
- Coleman A (2012) The effect of transport infrastructure on home production activity: evidence from rural New York, 1825–1845. Working Paper 12–01, Motu Economic and Public Policy Research, Wellington
- Coleman A, Grimes A (2010) Betterment taxes, capital gains and benefit cost ratios. *Econ Lett* 109(1):54–56
- Dixit A, Pindyck R (1994) *Investment under uncertainty*. Princeton University Press, Princeton
- Duflo E, Pande R (2007) Dams. *Q J Econ* 122(2):601–646
- Eddington P (2006) The Eddington transport study. Main report: transport's role in sustaining the UK's productivity and competitiveness. HM Treasury, London
- Florida R, Gulden T, Mellander C (2008) The rise of the mega-region. *Camb J Reg, Econ Soc* 1(3):459–476
- Forman C, Goldfarb A, Greenstein S (2012) The internet and local wages: a puzzle. *Am Econ Rev* 102(1):556–575
- Fujita M, Krugman P, Venables A (1999) *The spatial economy: cities, regions and international trade*. MIT Press, Cambridge, MA
- Glaeser E, Gottlieb J (2009) The wealth of cities: agglomeration economies and spatial equilibrium in the United States. *J Econ Lit* 47(4):983–1028
- Glaeser E, Gyourko J, Saks R (2005) Why have housing prices gone up? *Am Econ Rev* 95(2):329–333
- Gramlich E (1994) Infrastructure investment: a review essay. *J Econ Lit* 32(3):1176–1196
- Grimes A, Liang Y (2010) Bridge to somewhere: valuing Auckland's northern motorway extensions. *J Transport Econ Policy* 44(3):287–315
- Grimes A, Ren C, Stevens P (2012) The need for speed: impacts of internet connectivity on firm productivity. *J Product Anal* 37(2):187–201
- Guthrie G (2009) *Real options in theory and practice*. Oxford University Press, Oxford
- Haughwout A (2002) Public infrastructure investments, productivity and welfare in fixed geographic areas. *J Public Econ* 83(3):405–428
- Miller R, Lessard D (2008) Evolving strategy: risk management and the shaping of mega-projects. In: Priemus H, Flyvbjerg B, van Wee B (eds) *Decision-making on mega-projects: cost–benefit analysis, planning and innovation*. Edward Elgar, Cheltenham, pp 145–172, Chapter 8
- Morrison C, Schwartz A (1996) State infrastructure and productive performance. *Am Econ Rev* 86(5):1095–1111

- O'hUallachain B (1999) Patent places: size matters. *J Reg Sci* 39(4):613–636
- Overman H, Rice P, Venables A (2010) Economic linkages across space. *Reg Stud* 44(1):17–33
- Roback J (1982) Wages, rents, and the quality of life. *J Polit Econ* 90(6):1257–1278
- Rosen S (1979) Wage-based indexes of urban quality of life. In: Mieszkowsji P, Straszheim M (eds) *Current issues in urban economics*. Johns Hopkins University Press, Baltimore/London, pp 74–104
- Siegfried J, Zimbalist A (2000) The economics of sports facilities and their communities. *J Econ Perspect* 14(3):95–114
- Wu J, Gopinath M (2008) What causes spatial variations in economic development in the United States? *Am J Agric Econ* 90(2):392–408

Sandy Dall'erba and Irving Llamosas-Rosas

Contents

19.1	Introduction	354
19.2	Growth Theory and Regional Development Policy	355
19.3	Empirical Evidence and Lessons Learned in the EU and the USA	356
19.3.1	European Regional Policy	356
19.3.2	Regional Policy in the USA	358
19.4	Looking Ahead	361
19.4.1	Including Spatial Dependence and Reporting the Right Measurements	361
19.4.2	Measuring the Actual Investments, Not Proxies	365
19.4.3	Combining Different Strands of Theory and Techniques	366
19.4.4	Need to Develop Tools That Foster Communication Between Stakeholders and Academia	367
19.5	Conclusion	369
	References	370

Abstract

In spite of the ongoing efforts that several countries made into promoting a more balanced economic development within their territory, economic growth theory and even empirical evidence do not come to a unanimous conclusion on the efficiency of public intervention. As such, this chapter reviews the various

S. Dall'erba (✉)

Regional Economics And Spatial Modeling (REASM) Laboratory, University of Arizona,
Tucson, AZ, USA

e-mail: dallerba@email.arizona.edu

I. Llamosas-Rosas

Department of Economics, University of Arizona, Tucson, USA
e-mail: llamosas@email.arizona.edu

strands of the theoretical literature, analyzes the results of empirical estimations in Europe and in the USA where regional development policies are already well established, and provides recommendations for future research in this field.

19.1 Introduction

Why would a country finance public programs aiming at minimizing regional inequalities within its territory? The first reason is because it is not obvious that regional disparities will disappear by themselves. Globalization and decreasing transportation costs have led to a fragmentation of the production process and increasing agglomeration in a few places. This process is often reinforced by internal migration and foreign investments. The second argument which results naturally from the previous one is that persistent regional inequalities have raised concerns of national solidarity. It is based on the idea that citizens should be given the same opportunities in terms of access to public services, such as health and education, and to jobs no matter where they live in the country. Furthermore, because problems in accessing jobs are sometimes exacerbated by problems of social exclusions and discontent that may take a violent form, regional cohesion may act as a form of social cohesion. Policy intervention also has an efficiency goal by enhancing competition, boosting productivity, and international competitiveness. Programs aiming at removing barriers to internal trade, fostering the movement of the factors of production, and enhancing fair competition belong to this category.

In addition, while changes in the exchange rate have often been used to support the economy of a country, their impact on regional cohesion at the subnational level is not necessarily straightforward. Because of differences in their specialization, economic structure, and trade linkages with foreign partners, regions are not all equally sensitive to changes in the exchange rate, so that devaluation can actually increase regional inequalities. Finally, in absence of systematic adjustments through devaluation, local decision-makers may want to concentrate their efforts on a tax policy that attracts the mobile factors of production. Empirical evidence indicates that it can lead to a tax competition across regions, thus taking the form of a “race to the bottom” where mobile factors are offered lesser tax levels and the decrease in tax revenues is compensated by higher taxes on the immobile factors. As a result, regional development policies can be seen as an efficient tool to avoid a regional tax competition as well.

Yet, and despite the phenomena highlighted above, implementation and support for regional policies are not straightforward because of the important controversy about their efficiency. From a theoretical viewpoint, regional policies lead to different conclusions in terms of convergence according to the economic growth school of thought one focuses on (Dall'erba and Le Gallo 2008). From a practical perspective, several countries experience a lot of difficulties in assessing which regional development strategies work as empirical evidence shows conflicting results.

Therefore, Sect. 19.2 of this chapter is devoted to the theoretical impact of regional development policies under the lens of the neoclassical growth,

endogenous growth, and new economic geography theories. The lack of consensus in their predictions is corroborated by empirical evidence, as indicated in Sect. 19.3 which highlights the results of the key econometric studies that have measured the impact of public spending on regional growth. While many countries have ongoing regional development policies, we focus on the cases of the European Union and the USA. No other country or group of countries has a regional policy as developed and studied as the European Union. Starting with the first enlargement from 6 to 9 countries in 1973, it has since then been an intrinsic part of the EU integration process, and up to one-third of the European budget is devoted to it. Regarding the USA, the presence of a highly mobile labor force and of a federal tax adjustment mechanism have guaranteed that regional inequalities are much less pronounced than in Europe. However, the country is not immune of regional imbalances either. In addition, the implementation of the American Recovery and Reinvestment Act (ARRA) by the Obama administration in 2009 represents an example of why it is important to assess what regional development programs have been successful in the past in order to draw the right strategies for the future. As a consequence, the fourth section will provide a list, although not exhaustive, of elements that recent academic contributions have highlighted as necessary to consider for future developments in this field. Finally, the last section will provide a summary and concluding remarks.

19.2 Growth Theory and Regional Development Policy

From a theoretical viewpoint, the expected impact of regional policies is not straightforward. It varies from one strand of growth theory to the next. In a neoclassical framework based on Solow (1956), investments in physical capital per worker lead to a higher steady-state income. However, due to the decreasing marginal product of capital, the rate of investment must decline toward the steady-state income where the stock of capital per person is constant and of which growth is completely determined by technology. Therefore, regional policies in poor regions may stimulate their growth above their usual steady-state level, but it is only transitional and does not raise the steady-state income in the long run. On the other hand, public policies are granted a more central role in influencing long-run growth in the endogenous growth theory. Rejecting the neoclassical assumption of decreasing returns to scale, the endogenous approach sees public infrastructure as an input in the production function; hence, its presence increases the marginal product of private capital which fosters capital accumulation and growth. However, the addition of public capital in the production function does not allow one to look explicitly at the impact of regional policies on industry location. Indeed, firms choose to locate/relocate not only according to the transfers of purchasing power to the poor areas that accompany a regional policy, but also based on the effect of the latter on capital returns and trade costs between and within regions. Hence, the theoretical approach that is the most appropriate when analyzing the impact of regional policies on growth is the

new economic geography literature such as Fujita et al. (1999), as a large share of regional programs is often devoted to transportation infrastructure. Its appeal lies in returns which appear in the short run, a convenient feature for political purposes, and in its capacity to promote accessibility to/from any area which is commonly seen as beneficial to its economic development. The reality is more complicated. Supporting investments in interregional transportation infrastructure yields a decrease in transportation costs, which affects the process of industry location and often reinforces the agglomeration process which is already taking place in the rich regions. Agglomeration is due to the local labor/consumer market, knowledge externalities, local input–output linkages, and local infrastructures which are more developed in rich areas than poor ones.

Empirical evidence indicates that at the subnational level, transportation networks are firstly developed within and between rich regions because this is where the demand for transportation is the highest. In addition, in some cases, the transportation network is based on hub-and-spoke connections, like in Spain, where transportation costs from the hub (Madrid) to any spoke are lower than between spokes. As a result, connecting a poor area to the existing network may increase its accessibility, but the literature indicates that gains in accessibility will always be relatively higher in the central location than in the peripheral and poor one (Vickerman et al. 1999).

Even supporting transportation infrastructure projects within poor areas does not necessarily guarantees their sudden attractiveness as the spillovers they generate may be too small to counterbalance the agglomeration process already at work in the rich areas. In other words, accessibility is not the only challenge that poor areas have to deal with. It is often accompanied by a lack of infrastructure of any type, a less educated and smaller labor force and less efficient or nonexistent local input–output linkages. While it may not be a problem to sectors interested in access to natural resources or cheaper labor, assuming that interregional wage differences exist, the poorest areas often offer very few factors to promote location/relocation within their territory. As a result, several authors in the field of regional development have come to qualify regional policy as a trade-off between efficiency, which can be achieved by fostering agglomeration in the rich areas, and equity, meaning that public spending is used to maintain some level of economic activity and well-being in the poor places.

19.3 Empirical Evidence and Lessons Learned in the EU and the USA

19.3.1 European Regional Policy

Three main groups of results appear in the literature that estimates econometrically the effectiveness of the European cohesion policy on regional growth. Most of them focus on the so-called structural funds, the main tool of the EU regional policy. The first group, which concludes that structural funds have a significant and positive

impact, is composed of the studies of Fayolle and Lecuyer (2000), Cappelen et al. (2003), Beugelsdijk and Eijffinger (2005), and more recently Becker et al. (2010). The contribution by Fayolle and Lecuyer (2000) concludes that the regions that benefited the most from structural assistance are the wealthiest regions in the poorest countries. They explain that the reason for this is twofold: first, the new demand generated by structural funds support in the poor regions is supplied by the rich regions of the same country and, second, new transportation infrastructure helps the rich regions sell their products to the poor ones. Fayolle and Lecuyer (2000) are also the first ones to tackle the issue of co-funding which obliges the recipient regions to provide a share (between 15 % and 85 %, depending on the project) of the investment cost, a practice that softens the redistributive effects of the funds (Dall'erba and Le Gallo 2008). This level of detail allows Fayolle and Lecuyer to account for the actual amount of public spending in each regional economy, as structural funds per se are just a fraction of it.

The conclusions of Cappelen et al. (2003) are somewhat similar to the ones of Fayolle and Lecuyer (2000) as it also indicates that support is the most efficient when it is allocated to regions with a good economic environment, such as low unemployment and high R&D capabilities, which often are experienced in the most developed recipient regions. Hence, support is least efficient where it is most needed, which supports the idea that there is indeed an efficiency-equity trade-off. Finally, to our knowledge, the most recent contribution of the estimation of the impact of the funds is Becker et al. (2010). They focus explicitly on the group of regions recipient of objective 1 structural funds (allocated to regions of which per capita GDP is below 75 % of the EU average) and regions which qualify for these funds but did not receive them. They do not consider the amount of funding allocated to each region, but instead whether a region is recipient or not. Their approach differs from previous works since they focus on European regions of similar economic development level.

The second group of studies concludes that the impact of the funds is either nonsignificant or significant but negative impact. This group consists of the two studies by Dall'erba and Le Gallo (2008) and by Fagerberg and Verspagen (1996), respectively. Dall'erba and Le Gallo (2008) include a spatial econometric approach to convergence which allows them to account for the nonrandom distribution of structural funds and regional income; to proxy various variables at the origin of spillover effects, such as interregional trade, migration, technology externalities; and to measure coefficient estimates which are efficient. While their 2008 contribution pools all forms of structural funding together, the Dall'erba and Le Gallo (2007) work proposes an approach disaggregated by cohesion objective.

The third and final group of studies advocates for more mitigated conclusions on the impact of the funds. This group is composed of Rodriguez-Pose and Fratesi (2004), Ederveen et al. (2002), Ederveen et al. (2006), Dall'erba and Le Gallo (2007), Esposti and Bussoletti (2008), Bähr (2008), and Mohl and Hagen (2010). Focusing on objective 1 regions only, Rodriguez-Pose and Fratesi (2004) are the first ones to measure if the type of project financed, such as support for human capital or for agriculture, and the time it takes for funding to support growth (up to 7 years) matter.

They conclude that support to infrastructure and to businesses does not have a significant effect, even in the long run. On the other hand, investment in education and human capital has medium-term positive effects, while support to agriculture has short-term positive effects on growth. Dall'erba and Le Gallo (2007) are also driven by the desire to differentiate the impact of funding by the type of project this funding finances. In the absence of data for every project, they focus on all the categories of structural funds instead and pay attention to the amounts of additional funds, as in Fayolle and Lecuyer (2000). They find that peripheral regions are significantly but very lightly affected by some structural funds (objectives 1 and 3 & 4 funds, Community Initiatives), whether additional funds are accounted for or not. Based on a spatial econometric approach, they also highlight that peripheral regions seem more affected by the funds allocated to their neighbors than to themselves, more especially when they are objectives 2, 3 and 4 funds, or Community Initiatives. In the frame of a spatial panel setting, Mohl and Hagen (2010) concur that the conclusions are sensitive to the cohesion objective that is being analyzed.

Ederveen et al. (2002) claim that the results depend on whether convergence is measured without fixed effects (convergence across all the European regions) or with a national or regional fixed effect. They conclude that the more optimistic one is about convergence (no fixed effect), the less efficient structural funds spending appears to be, and vice versa. They are the first ones to highlight three detrimental mechanisms in the allocation of the funding across regions: rent seeking which takes place when regional governments design projects that meet the criteria of the EU but are not necessarily effective in stimulating growth; moral hazard which happens when local/regional authorities use EU funds for low-productive projects, so as to keep their region within the eligibility criterion for cohesion support; and crowding out which represents the fact that EU support creates a disincentive for local/regional/national governments to support their poor regional economies themselves. This substitution effect also takes place when EU funding reduces the incentive of the private sector to invest locally and/or workers to migrate to more productive areas, which would promote greater cohesion. Both Esposti and Bussoletti (2008) and Bähr (2008) find that structural funds per se have a negative impact on regional growth, but their impact becomes positive and significant when they interact with another variable such as R&D investments or human capital (in the case of Esposti and Bussoletti 2008) or decentralization, measured by the level of regional autonomy, in the case of Bähr (2008).

19.3.2 Regional Policy in the USA

In the USA, the earliest contribution focusing on the role of public capital on output is Aschauer (1989). His findings rely on national-level data and put the elasticity of public capital at 0.39. This result is somewhat similar to the one of Munnell (1990a) who obtains an elasticity of public capital (net of military spending) of which magnitude is between 0.31 and 0.37. However, both studies adopt a national-level approach. Since then, an increasing number, though still not very large, of studies

have focused its attention on the subnational level. For instance, Munnell (1990b) measures the participation of public capital among 48 states, assuming a Cobb-Douglas production function in levels. At the regional level, it is necessary to have an estimate of private and public capital stocks for each state, since the elasticity of the factors is measured by a Cobb-Douglas production function. As such, Munnell (1990b) develops a methodology that distributes the national stock of capital to each state. It allows her to find an elasticity of public capital on output of around 0.15 in the unconstrained equation and of magnitude in between 0.06 and 0.08 when the Cobb-Douglas coefficients are constrained by constant returns to scale (the sum of their elasticity equals one). Based on a panel data set for the 48 contiguous states over 1969–1986 and capital stock data from Munnell (1990b), Holtz-Eakin (1994) finds an elasticity of 0.203, which is in line with previous works. However, once he uses more complex estimation techniques, such as a fixed effect approach, IV and GLS, he does not find any significant effects.

More recently, Shioji (2001) measures the impact of public capital on economic growth based on a beta-convergence model and a set of panel data for the US states over 1973–1993. While he uses different econometric techniques to refine his results (GMM, LSDV), his approach does not include the role of private capital. His findings are ambiguous since he finds a range of impacts that goes from 0.572 (pooled regression) to 0.407 or even nonsignificant impact (based on LSDV and GMM). Once he disaggregates public capital into education spending and infrastructure spending, his results indicate a negative impact of the former and a positive one of the latter. The negative impact of public spending for education on growth is a result that several other US-focused studies have highlighted also. It indicates the countercyclical nature of this type of policy, and it reflects the high degree of mobility of US workers (Garcia-Milà et al. 1996). Paying attention to local effects, Nizalov and Loveridge (2005) measure the impact of economic development policies and highway infrastructure on growth and jobs across Michigan counties. They define three types of public expenditures: the Michigan Economic Growth Authority (MEGA), a program that grants businesses with tax credits for 8–20 years and targets investments and job creation; the Renaissance Zone (RZ) which provides local tax waivers to firms and individual residents of economically distressed areas; and the Brownfield Development Authority (BDA) which targets the redevelopment of blighted, functionally obsolete, and contaminated sites on Brownfield sites and highway infrastructure. Their approach is a linear estimation of the impact on growth of the above programs in addition to education, manufacturing, government, farming, and business concentration. They find ambiguous results, from a positive and significant effect of highways on job growth to a negative and significant effect of MEGA on income.

Two key studies on the 48 contiguous US states are Garcia-Mila and McGuire (1992) and Garcia-Milà et al. (1996) who focus on the impact of publicly provided inputs on income. In the former article, the authors use a Cobb-Douglas production function and a panel data set over 1969–1983. Public capital is split between highway capital (expenditures on highways by state and local governments) and support to education (state and local expenditures for K–12 and postsecondary education).

They find a positive and significant impact of highways on output (0.045) and a positive and significant impact of publicly provided education with a magnitude in between 0.165 (without a variable of median years of schooling) and 0.072 (with a variable of median years of schooling). When it comes to private capital, the estimated elasticity is in between 0.373 and 0.449 when it is measured as capital in equipment, and it is in between 0.027 and 0.104 when it is measured as capital structures. In the latter study (1996), they extend previous results by considering highways as well as water and sewers as publicly provided inputs over almost the same time period (1970–1983). They also use various fixed and random state effects to address the issue of heterogeneity in the data. Without controlling for state effects, they find similar results as Munnell (1990b) where highways as well as water and sewers have a positive and significant impact on output (0.37 and 0.069, respectively). However, once they controlled for state effects, the coefficients diminish to 0.120 for highways and 0.043 for water and sewers. Because they assume a potential serial correlation in their results, they run their model once more but on the variables measured into first differences. In this specification, all the publicly provided services appear to have a negative and significant impact, which is in tune with the conclusions of Holtz-Eakin (1994). Private capital has an impact in between 0.289 and 0.348, as usually found in the literature.

Using an extended version of Munnell (1990b) data, Lall and Yilmaz (2001) construct private and public capital stocks in order to estimate a beta-convergence model across the 48 contiguous states over 1969–1994 while controlling for business cycles by time-period dummies. Their results indicate a nonsignificant impact of lagged public capital in two specifications (without state or time dummies and with state dummies) and a significant but negative impact with state and time dummies, when the human capital variable is excluded from the equation.

Finally, the two most recent contributions on this topic have taken note of the theoretical advances advocated by the new economic geography literature as well as of the developments of the spatial econometric techniques to detect, model, and measure the presence of interregional externalities. As such, the work of Garrett et al. (2007) provides a spatial econometric estimation of beta-convergence across states over 1977–2002. Among the explanatory variables, government expenditures are measured as a proportion of state gross product, while local government revenues are captured by the share of state and local revenues. They find a negative and significant impact of government share of which magnitude (between -0.3097 and -0.3270) varies with the absence or presence of spillover effects. When it comes to the role of local revenues, their impact on growth is significant and negative with a range in between -0.0207 and -0.0218. Overall, they conclude that state-level fiscal policies can significantly influence income growth in neighboring states. The presence of interregional spillover effects is also at the core of the contribution of Dall'erba and Llamosas-Rosas (2012) who, in addition, measure *the actual* federal, state, and local public investments in education and other public capital from two databases: the Consolidated Federal Fund Reports (for federal spending) and the State and Local Government Finances (for local and state spending). It allows them to avoid using proxies for public

investments. Their results are in line with Holtz-Eakin (1994) since they find that public capital investments do not have a statistically significant impact, while public support for human capital has a negative and significant impact on per capita income. This corroborates the work of Kilkenny (2010) who shows that governments often neglect the negative feedback effects such as the rural “brain drain” of rural education to urban areas when rural development policies are implemented.

Looking at the previous results, we may wonder what reasons would explain such a diversity of outcomes both among European or US studies. We stipulate that there is a great deal of heterogeneity in the way they approach the same problem. The choice of the sample (only objective 1 regions vs. all the EU regions), time period (because of business cycles), estimation process (cross section vs. panel, presence or absence of fixed effects), the variables chosen (actual spending vs. some proxy), and the treatment of spatial dependence necessarily affect the estimation results. In addition, Ederveen et al. (2002) note that the conclusions are dependent upon the type of convergence estimated. In an absolute convergence framework, it is assumed that all the regions are converging to the same steady state, while adding spatial regimes (convergence clubs) or country dummies in the case of Europe allows for differences in regional steady states. The difference is not trivial since in the latter case the underlying assumption is that inequalities persist, even in the long run.

Differences in regional steady states are also controlled by the explanatory variables included in the model. The range and quality of explanatory variables that have been used in the studies above varies greatly. For instance, private investments statistics are available across EU regions but do not exist for the US states. As a result, they have to be constructed based on national data and following various methodologies such as the one of Munnell (1990b) or Garofalo and Yamarik (2002). There is no doubt that this affects the quality of the estimations.

19.4 Looking Ahead

19.4.1 Including Spatial Dependence and Reporting the Right Measurements

The last two decades have seen an increasing recognition of the role of spatial externalities in economic growth theory and empirical evidence. Because this movement has taken place in conjunction with a formalization of the spatial econometrics framework necessary for the estimation of various phenomena, the literature now displays a rather large number of studies estimating growth at the subnational level while accounting for spatial autocorrelation. As mentioned in Dall'erba and Le Gallo (2008), spatial autocorrelation refers to the fact that the spatial distribution of the variables used in the econometric model is not random. Rich areas tend to be close to other rich areas, and poor areas tend to be close to other poor areas. This phenomenon may come from factors such as trade, labor and capital mobility, technology, and knowledge diffusion that affect simultaneously nearby regions. It may also arise from model misspecifications

(omitted variables, measurement errors) or from a variety of measurement problems such as a mismatch between the administrative boundaries used to organize the data and the actual boundaries of the economic processes believed to generate growth.

If spatial autocorrelation proves to be present in an econometric model, the traditional assumption of independence of the error terms needs to be rejected; otherwise, it leads to unreliable estimates and inferences. Second, spatial autocorrelation allows the user to capture the presence of geographic spillover effects between observations, indicating that public funding in one location is not going to impact growth in the recipient location only. Third, spatial lags of the dependent variable can act as a proxy for omitted variables that are spatially dependent.

Among the studies listed in Sect. 19.3, only three have used these techniques in the European case and 3 in the US case. Details about the form of the spatial model they use, the definition of the variable of interest, and the estimated mean, minimum, and maximum impact appear in Table 19.1 above.

As can be seen in Table 19.1, different spatial models have been used in the literature to account for spatial dependence in a regression framework. More precisely, all these contributions have rejected the traditional (OLS) way of formulating an econometric growth model given by

$$y = \alpha \iota_n + X\beta + \varepsilon \text{ with } \varepsilon \sim N(0, \sigma^2 I_n)$$

where y is the growth rate of a period, X is a set of explanatory variables including public spending, and beta is a set of coefficients to be estimated. As such, the spatial econometric models they have measured in the contributions above are described below:

The spatial lag model (SAR):

$$y = \alpha \iota_n + \rho W y + X\beta + \varepsilon \text{ with } \varepsilon \sim N(0, \sigma^2 I_n)$$

$$y = (I_n - \rho W)^{-1} (\alpha \iota_n + X\beta + \varepsilon)$$

The spatial error model (SEM):

$$y = \alpha \iota_n + X\beta + u \text{ with } u = \rho W u + \varepsilon \text{ and } \varepsilon \sim N(0, \sigma^2 I_n)$$

$$u = (I_n - \rho W)^{-1} \varepsilon$$

$$y = \alpha \iota_n + X\beta + (I_n - \rho W)^{-1} \varepsilon$$

The spatial error and spatial lag model (SAC model):

$$y = \alpha \iota_n + \rho W_1 y + X\beta + u \text{ with } u = \theta W_2 u + \varepsilon \text{ and } \varepsilon \sim N(0, \sigma^2 I_n)$$

$$y = (I_n - \rho W_1)^{-1} (X\beta + \alpha \iota_n) + (I_n - \rho W_1)^{-1} (I_n - \theta W_2)^{-1} \varepsilon$$

Table 19.1 Summary of the impact of public spending on growth across spatial econometric studies

Primary study	Spatial model	Variable definition	Mean impact	Minimum impact	Maximum impact
European regions					
Mohl and Hagen (2010)	Spatial panel lag model	Structural funds per capita (objectives 1, 2, 3 – in log)	0.0003	-0.0092	0.0114
Dall'erba and Le Gallo (2008)	Spatial 2SLS lag model	Sum of structural funds per capita (in log)	-0.01 (not significant)	-0.01 (not significant)	0.002 (not significant)
Dall'erba and Le Gallo (2007)	Spatial error model	Sum of structural funds per capita (in log)	0.0005	-0.002	0.007
US states or counties					
Garrett et al. (2007)	Spatial lag and/or spatial error	Government share (first diff of log)	-0.3154	-0.3207 (spatial lag model)	-0.3097 (spatial lag and error model)
		Local revenue tax share (first diff of log)	-0.0211	-0.0218 (spatial error model)	-0.0207 (spatial lag and error model)
	Spatial lag model only	Government share (first diff of log)	-0.3169	-0.3214 (census divisions)	-0.3149 (census divisions)
		Local revenue tax share (first diff of log)	-0.0206	-0.0214 (census regions)	-0.0207 (census divisions)
Lall and Yilmaz (2001)	Lag on human capital only SLX	Public capital (constructed following Munnell 1990b)	0.002	-0.017 (public capital only)	0.036 (public and human capital) not significant
Dall'erba and Llamosas-Rosas (2012)	Spatial Durbin model	Public investment in infrastructure	-0.077	-0.154 (not significant)	-0.067 (not significant)
				Unrestricted model	Restricted model

The spatial cross regressive model (SLX):

$$y = \alpha I_n + X\beta_1 + WX\beta_2 + \varepsilon \text{ with } \varepsilon \sim N(0, \sigma^2 I_n)$$

The spatial Durbin model (SDM):

$$y = \rho W_1 y + \alpha I_n + X\beta_1 + W_2 X\beta_2 + \varepsilon \text{ with } \varepsilon \sim N(0, \sigma^2 I_n)$$

$$y = (I_n - \rho W_1)^{-1} (\alpha I_n + X\beta + W_2 X\beta_2 + \varepsilon)$$

where W is the spatial weight matrix and alpha is the constant term. Note that in the case of the SAC and SDM models, W_1 and W_2 can be equal or different. The reader can refer to LeSage and Pace (2009) for other forms of spatial models. It is important to understand that the interpretation of the parameters is not as simple as in a traditional linear regression model and may vary across spatial models. In the linear case, it is easy to interpret the impact on the dependent variable of a change in any explanatory variable. Indeed, the value of this impact is the magnitude of the coefficient. In addition, since the model assumes independence across observations, the effect of a change in any exogenous variable affects the dependent variable of that specific region only. Formally, $\partial y / \partial x^r = \beta_r$ (for any exogenous variable “ r ”). In a spatial econometric model, the presence of spillover effects often, but not always, makes the interpretation of the beta coefficients more complicated but richer. For instance, in a spatial error model as used in Dall'erba and Le Gallo (2007) or Garrett et al. (2007), the coefficient beta has the same meaning as in an OLS model. On the other hand, in an SLX model as used in Lall and Yilmaz (2001), a change in an explanatory variable is measured by

$$\partial y / \partial x^{r'} = (I_n \beta_r + W \theta_r)$$

In this formulation, since the weight matrix is standardized and contains zeros on the main diagonal, the coefficient β_r reflects direct effects, while θ_r captures local spatial spillovers. Finally, global spatial spillovers are measured in the frame of a spatial lag, as used in Dall'erba and Le Gallo (2008), Mohl and Hagen (2010), Garrett et al. (2007), or when estimating a spatial Durbin model as in Dall'erba and Llamosas-Rosas (2012). Indeed, in these studies the marginal effect is written as follows:

$$\text{SAR} : \partial y / \partial x^{r'} = (I_n - \rho W)^{-1} I_n \beta_r$$

$$\text{SDM} : \partial y / \partial x^{r'} = (I_n - \rho W)^{-1} (I_n \beta_r + W \theta_r)$$

Because the term $(I_n - \rho W)^{-1}$ can be expressed as the following infinite sequence: $(I_n + \rho W + \rho^2 W^2 + \rho^3 W^3 + \dots)$, it involves that a change in any region will affect not only the region itself, but its neighbors, the neighbors of its neighbors (which includes feedbacks to the original region), and so on.

It is important to stress that this nonlinear relationship implies that we cannot interpret the coefficients of a SAR or SDM model as in any of the other types of models, as the former measure both direct effects and spillover effects at the same time. This point has been overlooked in previous spatial econometric estimations of the role of public spending on growth where β is interpreted as in a nonspatial model, while the spatial lag coefficient is left capturing all kinds of spillover effects. Following the suggestion of LeSage and Pace (2009), the correct interpretation of a spatially lagged endogenous variable requires us to disaggregate the total effect of a change in an exogenous variable into direct effects which

capture the sum of the impacts in the region that experiences a change and indirect effects which capture the impact due to changes in other regions. In order to do so, direct effects are calculated as the mean of the main diagonal elements of the $n \times n$ matrices, while indirect effects correspond to the mean of the sum of the off-diagonal elements from each row of the $n \times n$ matrices. Details about the method to draw statistical inference on each effect can be found in LeSage and Pace (2009).

To our knowledge, there is no study so far that has used this approach to estimate the impact of regional development policies on growth in Europe or in the USA. The only study that comes close to it is Fischer (2011) who reports the magnitude of direct and indirect effects of (public and private) investments in physical and human capital across EU regions over 1995–2004.

19.4.2 Measuring the Actual Investments, Not Proxies

Earlier analyses, whether they focused on the European Union or on the USA, often used a (poor) proxy of the true amounts of public spending allocated across areas. Some popular proxies could be as simplistic as a dummy variable of which the value, 1 or 0, would reflect if a region is a recipient or not. The problem with this approach lies in its complete disregard for the actual amount of investment. Other contributions use the *stock* of human capital, such as the education level, in a Cobb-Douglas production function which is supposed to measure the role of human capital *investments*. Some authors recognize that the consistency and the reliability of their estimates, hence the quality of their conclusions, suffer from these proxies but not all.

While some scholars are to be blamed for their lack of rigor in collecting/constructing the appropriate data when they are available, the list of challenges they meet includes, but is not limited to the following: availability of the data at all, availability in electronic format vs. hard copy, updated vs. outdated data, data covering all types of projects vs. some only, detailed description of the project financed (for instance, “transportation” is too vague), description of the region where the projects have been allocated (the finer spatial scale the better), and data that correspond to the actual payments vs. investment commitments.

However, the prospects are much better. For instance, after years of relying on hard copy reports displaying data that would present many of the challenges listed above, the European Commission has moved on to creating a “computerized monitoring systems and electronic data exchange” site that serves as a unique reference for documenting the ways in which the funds are being used. The site is accessible here: http://ec.europa.eu/regional_policy/sources/exchange/exch_en.htm. It represents a significant step in the right direction, even though it is still very far from the level of transparency and detail that one experiences when working with the data of the US Census Bureau’s Consolidated Federal Fund Reports. They report electronically all types of federal spending, whether for regional policy purposes or not, for every county or smaller spatial units on

a yearly basis since 1993. These data are available here: <http://www.census.gov/govs/cfr/>. Recently, Dall'erba and Llamosas-Rosas (2012) have relied on this database to estimate the role of public spending on the regional economies of the USA in the frame of a Cobb-Douglas production function.

19.4.3 Combining Different Strands of Theory and Techniques

Many of the studies described in Sect. 19.3 rely on the famous neoclassical growth model initiated by Solow (1956) even though its underlying assumption of diminishing returns to capital and the eventual presence of Galton's fallacy have raised some doubts on its theoretical and empirical relevance. As a result, future works should consider theoretical models that mix different strands of the literature. In that sense, the contributions of Garrett et al. (2007) and Dall'erba and Le Gallo (2007, 2008) are innovative because they add the presence of interregional spillovers to a traditional neoclassical framework. Based on spatial econometric techniques, their results allow them to measure the extent to which structural funds impact not only the region where they are allocated but on neighboring regions as well. Another contribution that blends the various schools of economic growth theory even further is Ertur and Koch (2007). Based on a Cobb-Douglas framework, they propose to distinguish and model three factors that explain growth in technological progress: the first part is the stock of knowledge that is shared by all the firms and grows at an exogenous and constant rate as is usually assumed in the neoclassical approach. The second part of it is generated by the presence of knowledge externalities between nearby firms as described in the endogenous growth theory. Neither the first nor the second elements account for the role of dependence over space which has been brought to the forth by the new economic geography literature, which is why they attribute the third and final part of technological progress to localized interregional knowledge externalities.

Beyond a better integration of theoretical approaches, future contributions will also emphasize the need to integrate modeling techniques further. Spatial econometrics has become a popular and straightforward way to model and measure spatial dependence, but it does not rely on the “true” factors at the origin of interregional spillovers. It relies on a matrix of geographical proximity across regions which has the advantage of being determined exogenously. However, when it comes to economic growth and regional policy, it is mostly trade and migration that explains spatial dependence; hence, techniques capturing these interregional flows need to be adopted. It is the essence of interregional input–output (IO) analysis that has experienced increasing popularity since the early contributions of Wassily Leontieff. However, interregional IO data are long and costly to gather; hence, some authors such as Rey (2000) have suggested complementing the traditional input-out techniques with spatial econometrics to generate multiregional linkages that are both industrially and spatially disaggregated.

From a regional policy point of view, the advantage of combining techniques is twofold. First, it would allow scholars to avoid the “one-size-fits-all” approach that

has prevailed in the field of regional development. Indeed, in a global (econometric) approach which is the setting most empirical studies rely on, the coefficient associated to each variable corresponds to the average impact of the latter on the dependent variable across the entire sample. As a result, global econometric estimates could, for instance, reveal a significant impact of regional spending (from a statistical point of view) on the average regional growth rate, while in reality it is a nonsignificant impact that should be found in some localities and a positive or negative one elsewhere. The advantage of an approach that would be more disaggregated lies in its capacity to expose significant local and sectoral variations, which are masked by a global and aggregated approach. Secondly, most empirical estimations of the impact of regional policies have overlooked the potential endogeneity of regional spending. This problem comes from the fact that regional spending is mostly devoted to regions with a low per capita GDP, a measurement that is intrinsically part of the dependent variable, growth. Dall'erba and Le Gallo (2008) address this problem and use a set of instrumental variables since their Hausman test results reveal that structural funds are indeed endogenous. As a result, an integrated IO-spatial econometric approach as suggested by Rey (2000) could alleviate this problem since, by definition, the IO approach models and measures shocks that are either endogenous or exogenous to the system at hand.

19.4.4 Need to Develop Tools That Foster Communication Between Stakeholders and Academia

Preparation of material for dissemination in the public policy arena is not necessarily the main objective of many scholars in the field of regional development. However, the latter component is extremely important and undervalued as stakeholders are much more project driven and interested in the policy arena than academic scholars. In addition, the former often take decisions based on reports that are not produced by the latter. As a result, the first author of this chapter has recently concentrated his efforts on developing an Internet-based, free-of-charge, tool called the Regional Economic Impact Simulator. It transmits complicated theory and estimation techniques commonly used among regional scientists to an audience of specialists and nonspecialists. In addition, it can be used as a decision-support tool for localities willing to compare the returns of various kinds of investments.

Based on a webGIS platform, the Regional Economic Impact Simulator allows anyone to build a regional policy scenario of his/her choice and to visualize on a map, in a matter of seconds, how regional economic growth is modified as a result of it. Because of interregional interactions captured by spatial econometric means, it is not only the locality where the scenario is implemented that will experience a change in growth, but also the entire system.

An example of regional policy scenario is depicted in Fig. 19.1 below. It reflects how economic growth over 2000–2008 in each of the counties of the sample has been

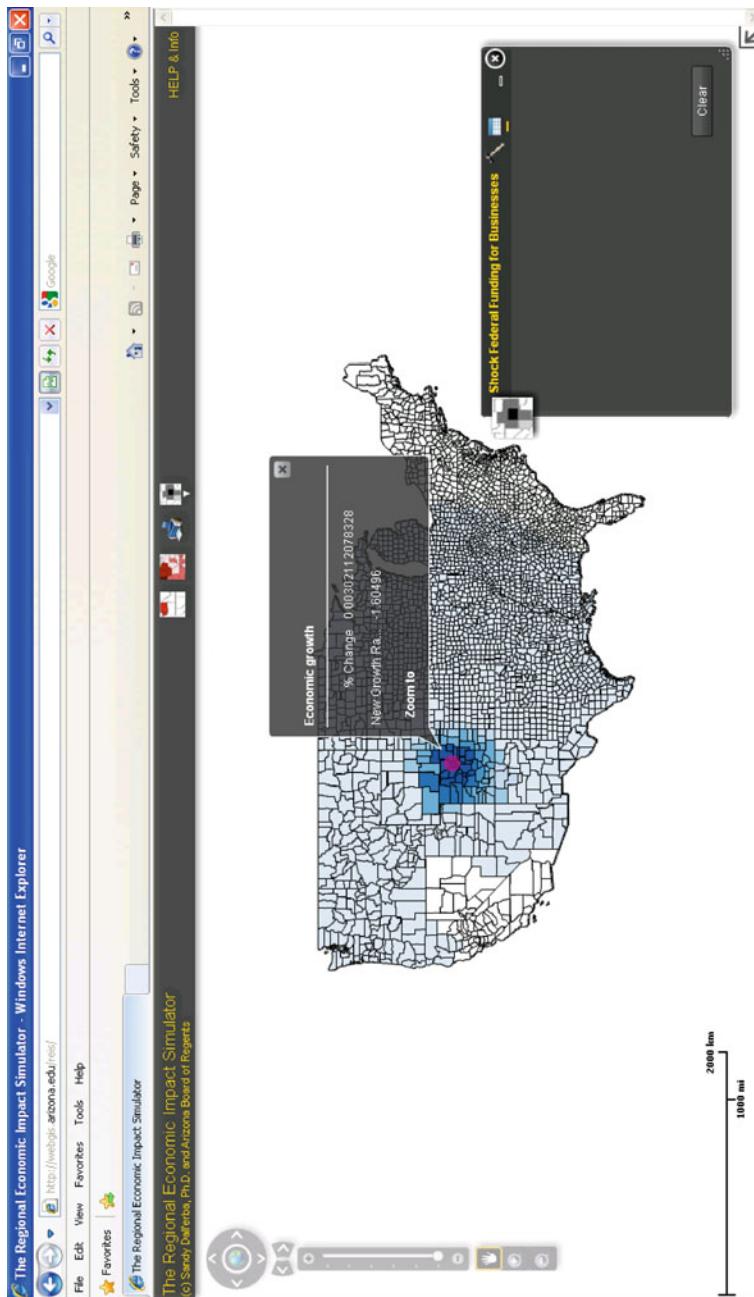


Fig. 19.1 Changes in county-level economic growth resulting from a simulated 50 % increase in federal support for private businesses in Boulder, Colorado

modified as a result of a 50 % increase in federal support for private businesses in Boulder, Colorado, over the same period. While this county is the one which has experienced the greatest change in growth, growth has also spread to its neighbors and the neighbors of its neighbors, etc., but with a decreasing magnitude as distance from Boulder increases. Note also how a pop-up window allows users to get details about the magnitude of the change experienced in each county. Many more simulations can be performed on the Regional Economic Impact Simulator as users are given the freedom of implementing shocks of any magnitude, on any of the 16 explanatory variables used in the model and any of the 3,076 counties of the database. The Regional Economic Impact Simulator is accessible here: <http://webgis.arizona.edu/reis/>. While it is, to our knowledge, the only tool of this nature available at the moment, we anticipate that the increasing desire for transparency and accountability in the use of public funding will lead to many more free, Internet-based, decision-support tools.

19.5 Conclusion

Increasing interest for regional development policies has led to new theoretical advances and a growing number of empirical works, but it has not succeeded in providing a standard model for economic development intervention. Past empirical evidence indicates that the conclusions are very sensitive to a set of parameters, such as sample size, estimation method, time period, and quality of the variables used, which confuses academic scholars, stakeholders, and policy-makers. An example would be the 2009 American Recovery and Reinvestment Act which led 200 economists to predict that it would benefit the US economy and 200 others to forecast the exact opposite. As a result, uncertainty calls for “place-tailored” policies where intervention is designed to meet the specific economic and geographic characteristics of the recipient locality targeted for development. For instance, it is now well accepted that transportation infrastructures are not necessarily an efficient tool to promote equality as their impact does not always benefit the recipient area. Indeed, they may lead to depopulation of the latter and increase agglomeration in rich regions.

The most recent contributions in the field of regional development are focusing their efforts on avoiding some of the shortfalls of the past such as the correct interpretation of spatial models and the overwhelming reliance on proxies as opposed to the actual amounts of public spending. In addition, there is an increasing desire to combine the different strands of economic growth theory with each other while integrating further the set of regional science techniques already available. Ultimately, it should provide a more complete picture of the regional dynamics at stake and of the actual role of policy intervention. Last but not least, regional development practitioners need to rely more often on current technology, such as webGIS, to bridge the gap between the interests of stakeholder and the expertise of academic scholars as well as demonstrate to the general public the level of transparency and accountability they operate in. Only an increasing awareness of today’s regional development challenges will oblige us all to make future interventions more effective and efficient than past ones.

Acknowledgment We would like to thank the National Science Foundation (Program NSF-GSS PD-1352) for providing financial support.

References

- Aschauer D (1989) Is public infrastructure productive? *J Monet Econ* 23(2):177–200
- Bähr C (2008) How does sub-national autonomy affect the effectiveness of structural funds? *Kyklos* 61(1):3–18
- Becker SO, Egger P, von Ehrlich M (2010) Going NUTS: the effect of EU structural funds on regional performance. *J Public Econ* 94(9–10):578–590
- Beugelsdijk M, Eijffinger S (2005) The effectiveness of structural policy in the European Union: an empirical analysis for the EU-15 in 1995–2001. *J Common Mark Stud* 43(1):37–51
- Cappelen A, Castellacci F, Fagerberg J, Verspagen B (2003) The impact of EU regional support on growth and convergence in the European Union. *J Common Mark Stud* 41(4):621–644
- Dall'erba S, Le Gallo J (2007) The impact of EU regional support on growth and employment. *Czech J Econ Financ* 57(7–8):325–340
- Dall'erba S, Le Gallo J (2008) Regional convergence and the impact of structural funds over 1989–1999: a spatial econometric analysis. *Pap Reg Sci* 87(2):219–244
- Dall'erba S, Llamosas-Rosas I (2012) The impact of private, public and human capital on the US States economies: theory, extensions and evidence. In: Karlsson C, Andersson M (eds) *Handbook of research methods and applications in economic geography*. Edward Elgar, London, under review
- Ederveen S, Gorter J, de Mooij R, Nahuis R (2002) Funds and games: the economics of European cohesion policy. CPB Working Paper, pp 1–103
- Ederveen S, de Groot HLF, Nahuis R (2006) Fertile soil for structural funds? A panel data analysis of the conditional effectiveness of European cohesion policy. *Kyklos* 59(1):17–42
- Ertur C, Koch W (2007) Growth, technological interdependence and spatial externalities: theory and evidence. *J Appl Econ* 22(6):1033–1062
- Esposti R, Bussoletti S (2008) Impact of objective 1 funds on regional growth convergence in the European Union: a panel-data approach. *Reg Stud* 42(2):159–173
- Fagerberg J, Verspagen B (1996) Heading for divergence? Regional growth in Europe reconsidered. *J Common Mark Stud* 34(3):431–448
- Payolle J, Lecuyer A (2000) Regional growth, national membership and European structural funds: an empirical appraisal. *La Rev l'OFCE* 2:1–31
- Fischer M (2011) A spatial Mankiw-Romer-Weil model: theory and evidence. *Ann Reg Sci* 47(2):419–436
- Fujita M, Krugman P, Venables AJ (1999) The spatial economy. MIT Press, Cambridge
- Garcia-Milà T, McGuire TJ (1992) The contribution of publicly provided inputs to states' economies. *Reg Sci Urban Econ* 22(2):229–241
- Garcia-Milà T, McGuire TJ, Porter RH (1996) The effect of public capital in state-level production function reconsidered. *Rev Econ Stat* 78(1):177–180
- Garofalo GA, Yamarik S (2002) Regional convergence: evidence from a new state-by-state capital stock series. *Rev Econ Stat* 84(2):316–323
- Garrett TA, Wagner GA, Wheelock DC (2007) Regional disparities in the spatial correlation of state income growth, 1977–2002. *Ann Reg Sci* 41(3):601–618
- Holtz-Eakin D (1994) Public-sector capital and the productivity puzzle. *Rev Econ Stat* 76(1):12–21
- Kilkenny M (2010) Urban/regional economics and rural development. *J Reg Sci* 50(1):449–470
- Lall SV, Yilmaz S (2001) Regional economic convergence: do policy instruments make a difference? *Ann Reg Sci* 35(1):153–166
- LeSage J, Pace K (2009) Introduction to spatial econometrics. CRC Press, Boca Raton

- Mohl P, Hagen T (2010) Do EU structural funds promote regional growth? New evidence from various panel data approaches. *Reg Sci Urban Econ* 40(5):353–365
- Munnell AH (1990a) Why has productivity growth declined? Productivity and public investment. *N Engl Econ Rev* January/February:3–22
- Munnell AH (1990b) How does public infrastructure affect regional economic performance? *N Engl Econ Rev* September/October:11–32
- Nizalov D, Loveridge S (2005) Regional policies and economic growth: one size does not fit all. *Rev Reg Stud* 35(3):266–290
- Rey S (2000) Integrated regional econometric + input–output modeling: issues and opportunities. *Pap Reg Sci* 79(3):271–292
- Rodriguez-Pose A, Fratesi U (2004) Between development and social policies: the impact of European structural funds in objective 1 regions. *Reg Stud* 38(1):97–113
- Shioji E (2001) Public capital and economic growth: a convergence approach. *J Econ Growth* 6(3):205–227
- Solow R (1956) A contribution to the theory of economic growth. *Q J Econ* 70(1):65–94
- Vickerman R, Spiekermann K, Wegener M (1999) Accessibility and economic development in Europe. *Reg Stud* 33(1):1–15

Section III

Innovation and Regional Economic Development

Knowledge innovation firms networks development local firms proximity spatial activities product
information economies diffusion districts
ideas
cities distance
learning change
global technology urban flows products
agents social region geography sources agglomeration
growth industries market interaction milieu clusters

Edward J. Malecki

Contents

20.1	Introduction	376
20.2	The Standard Model: Innovation as Knowledge Production	376
20.3	What Is Innovation?	377
20.4	Two Faces: The Multidimensional Nature of Innovation	378
20.5	How Innovation Works: Linear or Complex?	380
20.5.1	Entrepreneurship as Innovation	380
20.5.2	Innovation as a Complex Process	381
20.6	Global R&D	382
20.6.1	Fact 1: Innovation Is Dispersing Globally	383
20.6.2	Fact 2: Places Are Competing for Talent and Brains	383
20.6.3	Keeping Track	384
20.7	Policy: Changing and Reacting to Changes in the Geography of Innovation	385
20.7.1	Regional Innovation Policy: Constructing Advantage	385
20.7.2	Policy in a World of Global Production Networks and Global Value Chains	387
20.8	Conclusions	387
	References	388

Abstract

This chapter surveys the topic of the geography of innovation – not the economics of innovation – and asks several questions: What is innovation? Who innovates? Where do they learn to innovate? The research focus has shifted from innovation and technology to the broader issues of knowledge and innovative capability. The empirical literature has been much narrower in scope, previously focusing on research and development (R&D) and now rarely looking beyond patents.

E.J. Malecki

Department of Geography, Ohio State University 1036 Derby Hall, Columbus, OH, USA
e-mail: malecki.4@osu.edu

The chapter surveys a broader set of innovation indicators – inputs, outputs, and hidden innovation, much of which is uncovered in large-scale surveys. Empirically, there is a global shift in innovative capability toward Asia, primarily in R&D (but less so in basic research) and in process innovation related to manufacturing. The overall pattern is one of persistent spatial concentration. As a result, a thriving business has emerged to craft policies to enhance innovation and to “construct advantage” in an uncertain competitive landscape. Finally, the actors in innovation include not only individual scientists and inventors but also the organizations that employ them, such as universities and firms. It is entrepreneurs who largely determine how innovation is exploited. The fruitful concept of the knowledge filter and the role of entrepreneurship and the geography of entrepreneurship provide clues to the patterns seen.

20.1 Introduction

Innovation is fundamental to economic growth and to variations in economic development across space. Innovation is a dynamic process – bringing about creative destruction and shifting the locus of innovation across industries and among locations. Innovation is a broader concept than technical (or technological) change alone; it includes ideas and knowledge that precede actual innovation, the learning that takes place through experience, and the synthesis of knowledge and complementary assets to utilize them profitably. Our understanding of the geography of innovation, however, has become too narrowly conceived as the topic has attracted a flurry of research attention from economists.

This chapter reviews what we know from the work of economists as well as from geographers. It begins with the simple world as grasped by patent data and models based on those data. Second, it examines what we know about innovation as a dynamic and complex – even messy – process. Third, this dynamic, complex messiness is seen in the location of innovation and of innovative capability: the geography of innovation at the global scale, which is itself the outcome of several distinct flows and forces. Fourth, the chapter reviews briefly the degree to which policy can influence the geography of innovation.

20.2 The Standard Model: Innovation as Knowledge Production

The standard economic view includes a knowledge production function, in which innovative output, typically measured as patents, results from innovative inputs, specifically research and development (R&D) by firms. Within regions, knowledge spills over from universities and industrial R&D, and these spillovers decline with distance. Griliches (1990: 1669) acknowledged “a whole host of problems” with patent data: “Not all inventions are patentable, not all inventions are patented, and the inventions that are patented differ greatly in ‘quality,’ in the magnitude of

inventive output associated with them". However, most researchers have been persuaded by his conclusion that "[i]n spite of all the difficulties, patent statistics remain a unique resource for the analysis of the process of technical change. Nothing else even comes close in the quantity of available data, accessibility, and . . . detail" (p. 1702). Consequently, there has been a flood of research using patents as the principal – and often the only – measure of innovation and its geography.

Reviewing the literature since Griliches (1990), Nagaoka et al. (2010) conclude that patent data alone are not sufficient for understanding the mechanisms of either a knowledge production function or knowledge spillovers. Too many important flows are not captured by patents, including other means of appropriability, inventions not patented, and patented inventions not used. More common means of firms to protect their innovations include secrecy, lead time, and complementary capabilities (Nagaoka et al. 2010).

Notably, other knowledge – particularly tacit knowledge – is ignored or assumed to spread as seen in patent citations. The importance of tacit knowledge is that it is not readily transferred, and therefore, transfers are difficult unless the parties are colocated (permanently or temporarily) in a locality (Asheim and Gertler 2005). The growth of publications by firms also can be seen as an attempt both to find new access to external knowledge and to signal the existence of tacit knowledge and other unpublished resources. Publishing allows a firm's researchers to become involved or take part in academic activities, with access to the epistemic communities of researchers. In return, the firm expects access to the tacit knowledge of academics in the field.

Not all industries are the same. High-technology, or high-tech, industries have received a great deal of attention, and they have informed our knowledge of how innovation occurs in those sectors. Innovation is managed differently across industries: fast-changing industries may be more creative but also less efficient, whereas slow-changing industries emphasize efficiency over creativity.

Patenting is not typical in all sectors; the pharmaceutical industry is particularly dependent on patent protection. The glamour of biotechnology and blockbuster drugs has led to an overemphasis on the importance of patenting. We know a great deal about biotechnology, in part, because the industry fits the R&D-based model of prevailing theory and, in part, because it is relatively small and localized in few locations. We know that patents generally are highly concentrated in large cities.

Those who study innovation at the scale of the firm, rather than the region, have noted the evolution from the development of something new to a process of creativity to a process dependent on knowledge. A similar evolution has taken place in the management of R&D within firms.

20.3 What Is Innovation?

As a result of the broadening of definitions and of ways to measure innovation, broader views of innovation have become more widespread. They have developed largely within Europe, particularly in the context of Organisation for Economic Co-operation and Development (OECD) and European Union (EU) policy documents

(Mytelka and Smith 2002). The OECD definition of an innovation includes more than merely a new or significantly improved product (good or service) or process, but also the implementation of a new marketing method, or a new organizational method in business practices, workplace organization, or external relations.

Products and processes – some of which are patented – have been the object of research on innovation for decades; services, marketing methods, and organizational methods are more recent. The series of Community Innovation Surveys conducted in (an expanding number of) EU member states have added to our knowledge of how firms innovate. Similar surveys are now conducted in several countries outside Europe, with the United States conspicuously not among them.

To this list can be added *soft innovation* – new products offering aesthetic rather than functional appeal as well as those goods and services with a distinctly intellectual appeal (including books, films, art, or computer games). Soft innovation builds on the increased importance of aesthetic content in products. Aesthetic improvements, the outcome of soft innovations, are a principal source of product differentiation but generally cannot be patented. “Whereas patents require novelty and copyright requires originality, the counterpart for a trademark is distinctiveness. . . . Whereas patents are not available for aesthetic innovations, such innovations may be trademarked” (Stoneman 2010: 262). In addition, copyright protects material such as literature, art, music, sound recordings, films, and broadcasts, and design rights (or design patents) protect the appearance or visual appeal of products (Stoneman 2010). Trademarks and copyrights remain underappreciated and understudied aspects of innovation.

Many have focused on tacit knowledge, implicitly in tune with the idea of technologies as recipes. However, a great deal of tacit knowledge is needed beyond the technological procedures in any codified recipe. Tacit knowledge flows through many channels, such as the multitude of interactions and knowledge flows between economic entities such as firms (customers, suppliers, competitors), research organizations (universities, other public and private research institutions), and public agencies (technology transfer centers, development agencies) (Asheim and Gertler 2005). Organizational means for absorbing, integrating, and transforming knowledge have been a major focus of research.

Just as inputs beyond R&D (and of outputs beyond patents) are important in innovation, more than a patent is necessary for a firm to appropriate and profit from the gains from an invention. Even an imitator can outperform an innovator if the imitator has assembled a better set of critical complementary assets. It also is the case that R&D has purposes beyond only patentable innovations, on which more below.

20.4 Two Faces: The Multidimensional Nature of Innovation

Binaries and dichotomies are simple solutions to complex problems. Dichotomies are found in studies of knowledge and innovation, shedding light but also obscuring the actual workings of knowledge production and innovation. Knowledge is more complex than merely codified and tacit.

One of the more useful binaries is the observation that R&D has two faces; that is, firms invest in R&D not only to generate innovations but also to learn from competitors and knowledge sources outside the industry, such as university and government labs (Cohen and Levinthal 1989). To a large degree, subsequent research has followed either one path or the other. The first flood of research has focused on patents and linkages between patents through citations.

A second body of research, largely unconnected, has focused on learning by people and firms and on types of knowledge. Not all knowledge is used (nor perhaps even useful) immediately and is retained and accumulated for future use. This store of knowledge is to the individual's – and the firm's – absorptive capacity. Firms differ widely in their knowledge search strategies, with impacts on their absorptive capacities.

Other motivations for R&D beyond the need to build an absorptive capacity include the following: an intention to maintain the firm on the technological frontier, the search for reputation, building and signaling its competences, and entrance to networks, which are among the main incentives for firms to invest in R&D.

Systems integration, like R&D, also has two faces: the internal activities of firms as they develop and integrate the inputs they need to produce new products and services and the external activities of firms as they integrate components, skills, and knowledge from other organizations to produce ever more complex products and services. Complex systems of technologies (e.g., automobiles, aerospace systems, iPhones) require the integration of knowledge from many sources – technological as well as geographical. Systems integration and the skills required to translate and interpret across disciplines, jargon, language, and technologies – and to synthesize these into forms and routines usable within the organization – are neither easy nor straightforward.

Knowledge as created and used is not identical but differentiated. Three types of knowledge bases have been outlined by Asheim et al. (2011): *analytical* (science based), *synthetic* (engineering based), and *symbolic* (arts based). Alternative typologies are based on whether knowledge is codified or poorly articulated, spillovers are intentional or unintentional, and incentives to reveal and to capture knowledge are strong or weak. Once again, biotechnology (with few other industries) stands out as unusual.

A key debate in the literature concerns whether specialization or diversity within an agglomeration is most beneficial for spillovers. The consensus had begun to shift toward diversity, but the current consensus is swayed by research which shows that *related variety* is best (Asheim et al. 2011; Boschma and Frenken in Cooke et al. 2011; Iammarino in Cooke et al. 2011).

A recent addition to the roster of binaries is the distinction between local and nonlocal (or extralocal) knowledge sources. The impact of local “buzz” is particularly important in the creative and cultural industries, where symbolic knowledge, performance, and events perhaps outweigh the importance of codified, cumulative knowledge. A useful literature has grown on types of proximity – not only geographical, but also organizational, cultural, technological, cognitive, institutional, and social (► Chap. 26, “*Networks in the Innovation Process*” by Tranos). The summary by Moodysson and Jonsson (2007: 15) concerning Swedish biotechnology firms is appropriate more widely: “The convenience of local collaboration can never replace the extreme requirements of specialized knowledge, which forces them to seek collaborators on a global arena.”

Nuanced views deconstruct the meaning of proximity to an even greater extent: However, proximity has a different influence depending on the size of the city. The concept of *temporary clusters* also reflects the complexity of proximity: it need not be fixed in place or permanent to be beneficial. Crevoisier and Jeannerat (2009) develop a richer framework to understand the simultaneous need for both local and nonlocal sources. “The concept of ‘from elsewhere’ is now differentiated: the places are clearly identified, as are complementary and/or competing ones” (p. 1235). In other words, the globalization of knowledge does not reflect an amorphous “elsewhere.” It reflects known places where specialized knowledge is as great, or greater, than in a given locality.

Clusters do not necessarily have links to knowledge pools elsewhere (Vale 2011). Even when a region’s knowledge networks include pipelines to distant knowledge, that knowledge needs to be “anchored” and integrated with the regional knowledge base (Crevoisier and Jeannerat 2009; Vale 2011).

20.5 How Innovation Works: Linear or Complex?

The bottom line is that innovation flows from knowledge and ideas, broadly viewed, rather than from R&D, even if broadly defined. This point is implicit in the new growth theory, which results in positive-sum growth based on ideas. The recent criticism by Steinmueller (2010: 1190) is that “unlike the old growth theory which produced a central result, the ‘golden rule’ of accumulation, the new growth theory is still evolving” with considerable variety in its outcomes. The broader conception of knowledge and innovation is more explicit in the recent shift seen from innovation systems to knowledge systems.

Early work was based, explicitly or implicitly, on the *linear model*, which postulates that innovation begins with basic research, adding applied research and development (all still under the umbrella of R&D), followed by production and diffusion. A standard of innovation studies for decades, the linear model captures the temporal sequence of activities in innovation, is easily monitored in data gathering and appeals to policymakers because of its simplicity and logic. The linear model fits biotech and other science-based industries because of a key feature of the linear model – linearity – the fact that not everything occurs simultaneously. Regional analyses based on the linear model and on patent data are still common.

20.5.1 Entrepreneurship as Innovation

The linear model has been extended to encompass entrepreneurship. Ideas lead not only to new products and services but also to new firms and, in some cases, to clusters of new firms in new industries. Indeed, research and policy interest in entrepreneurship grew largely out of interest in technology-based clusters (Mason 2008).

The entrepreneurial process within the innovation process is captured best by the *knowledge filter*, a key element in the knowledge spillover theory of

entrepreneurship (Audretsch and Keilbach 2007; Acs et al. 2009). The knowledge filter is the set of barriers to converting research into commercialized knowledge.

The knowledge filter for academic research is (besides the challenge of converting basic science into applied knowledge) largely an institutional filter. It consists of organizational barriers, university policies, attitudes among faculty and university administrators against commercialization of research, and lack of incentives to pursue commercialization. Additional barriers within the academic knowledge filter reflect an inability to convert inventions into intellectual property (primarily in the form of patents) and to commercialize that intellectual property through licenses and start-ups.

Similar filters for industrial R&D reflect the difficulty in business organizations to convert research into intellectual property and to commercialize new products. Entrepreneurs are able to see a path and to assemble the networks necessary for commercialization. Many types of interfirm networks are needed, ranging from global to local and from formal links to informal networking (Lawton Smith 2008; Giuliani in Cooke et al. 2011).

20.5.2 Innovation as a Complex Process

What is missing from the dominant flow of the linear model are the feedbacks and interactions that are so crucial to innovation. Many interactions are contained within national boundaries or within regions (Asheim and Gertler 2005). These national and regional innovation systems are thought to largely define the institutions, cultures, and path-dependent strengths (and weaknesses) that vary from place to place. The idea of innovation as – and within – systems recognizes innovation as a complex and systemic phenomenon. Research on innovation systems also reflects this broad synthetic perspective (Fischer et al. 2001; Mytelka and Smith 2002; Soete et al. 2010; Wolfe in Cooke et al. 2011; ► Chap. 24, “Systems of Innovation and the Learning Region” by Cooke).

The spatial complexity of RISs and the operational complexity of learning have already pushed the linear model into the background. Caraça et al. (2009) suggest a multichannel interactive learning model that captures the complex flows and interactions among actors.

Work on innovation systems was at first national and technological. Subsequent research added sectoral systems and regional systems (Tödtling and Trippl in Cooke et al. 2011). Research has begun to recognize the overlap and boundary relations between national, sectoral, and technology-specific innovation systems and between technological systems and sectoral systems of innovation. Crossing international boundaries highlights the distinctiveness of each national innovation system as nations compete to stay innovative and thereby wealthy. Regional systems link to those in other national systems, thereby forming international innovation systems (Crevoisier and Jeannerat 2009; Soete et al. 2010).

Regional innovation systems (RISs) have attracted the research attention of economic geographers and regional scientists (Asheim and Gertler 2005). The actual

workings, activities, and policies, as well as measurements of contacts and linkages, are sometimes easier to grasp within a regional or local context than at the scale of the nation. Several terms are used to describe such local territorial innovation systems, such as clusters, territorial production complexes, productive systems, territorial systems, milieus, and local systems (De Propris and Crevoisier in Cooke et al. 2011).

The significance of RISs is that they represent “spatial knowledge monopolies” that attract investment and participation by transnational corporations (TNCs) (Cooke 2005). The central feature of the national and regional innovation systems is that while R&D activity still matters greatly, it is only one part of a larger system that includes education, training, government support, and linkages among sectors. Recent research suggests that regional clustering and networking (such as those found in innovative milieus) are less important than localized capacities to build global connections.

Intermediaries also can bring external knowledge to potential users. Knowledge-intensive business services (KIBS) are a particularly important source of innovative knowledge. Geographically, KIBS are highly concentrated at the top of the urban hierarchy.

Feldman and Kogler's (2010) eight stylized facts in the geography of innovation focus on the importance of proximity and location to innovative activity. Although the geography of innovation comprises agglomeration and spillovers (Feldman and Kogler 2010; ► Chap. 22, “Knowledge Flows, Knowledge Externalities, and Regional Economic Development” by Karlsson and Gråsjö), it is also much more. Agglomeration or clustering alone does not provide the ingredients within RISs necessary for collective learning – institutions, social capital, and entrepreneurs (Capello in Cooke et al. 2011). The advantages of agglomeration are well established, providing opportunities for sharing, matching, and learning. In general, large urban areas are expected, ceteris paribus, to have higher proportions of skilled workers, higher rates of innovation, and more rapid adoption of innovations, smaller places. However, all large and/or dense cities are not alike; they vary widely in culture and in institutional infrastructure.

What is left out, of course, is the complex of social dynamics, captured in part by the concept of social capital, which is fundamental to the cohesion (or lack of it) in a community. Power – especially the power exerted by TNCs – is key to the actual dynamics in many regions, but is omitted from most analyses of RISs.

The ground-up, largely local view of how the geography of innovation is constructed is primarily an economic view rather than a bird's-eye look at the changing geography of innovation. The following proceeds from the opposite direction: from the global to the local.

20.6 Global R&D

That R&D is global has been evident for over two decades. Through the 1990s, however, global R&D was largely triadic – distributed among (western) Europe, Japan, and North America. Since 2000, the “global landscape” of R&D has changed dramatically, reflecting major innovative effort in Asia outside Japan. The current

situation is a *global innovation system*, in which India, China, and the United States have leveraged the growing internationalization of innovation to offset weaknesses in their own national innovation systems.

The following focuses on two central shifts at the global scale: the location of innovative activity and the competition for talent. These shifts and the measures of them are tracked by a bewildering array of scoreboards of indicators.

20.6.1 Fact 1: Innovation Is Dispersing Globally

The geography of innovation used to follow the product cycle in a predictable manner, flowing from R&D, conducted only in high-income countries. The activities of TNCs and their global production networks have altered but not eliminated product cycle as an important concept at the global scale (Tichy in Cooke et al. 2011). The benefits of agglomeration economies appear to be greatest at the “birth” of new firms and diminish during the later stages of the industry life cycle.

During the 1970s – that is, before the rise of China – R&D had begun to globalize, becoming much more so during the 1990s to exploit sources of knowledge at the locations of customers and competitors. However, in-house R&D alone is no longer sufficient for a firm to be technologically competitive. In-house R&D must be complemented by external sources of innovation, which then need to be integrated into the firm’s structures and competences. These trends are captured in firms’ utilization of open innovation and the phenomenon of the double network.

The global innovative activities of TNCs are one force behind the shift from R&D being located only (or primarily) in rich countries. Another force is active efforts by firms – many state owned – in emerging economies to serve their own growing consumer markets. This means that in any industry, the number of pipelines a firm must maintain is increasing. As Crevoisier and Jeannerat (2009) stress, there are many knowledge sources, and links to them require effort to maintain rapport and productive contact. In short, research has increasingly become a borderless activity.

20.6.2 Fact 2: Places Are Competing for Talent and Brains

The global geography of innovation has been transformed primarily by the globalization of scientific and engineering talent, which Freeman (2010) suggests has proceeded rapidly along five related tracks. These are:

- Expansion of mass higher education worldwide
- Growth in number of international students
- Migration
- Non-immigration trips by academic visitors and conference attendees
- A rapid rise in international coauthorship and co-patenting

As these five changes have occurred, changes in national capabilities have taken place.

Migration is an important channel for the movement and spread of knowledge. Migration need not be considered a brain drain, but can be a *brain recirculation* as migration is less frequently permanent and as people construct multilocational careers and professional and personal lives. The term brain drain has been replaced by global competition for talent. In this ongoing competition or race for talent – and for highly skilled migrants – countries have implemented *competitive immigration regimes* as a new form of interjurisdictional competition.

There are benefits from such policies, seen the movement of the world's productive researchers toward nations with research infrastructure and strong R&D support.

At the level of policy, as opposed to theory, the geography of innovation primarily means the visible shifts in innovative capability, inputs, and outputs on the ground – the changing landscape of innovation. The new, more global pattern reflects the growing role of knowledge in the global economy, seen primarily in investments in tertiary education and R&D outside the OECD countries. Scientific publications are the result of creative efforts of people working in universities, government research institutes, and the R&D labs of private firms. The map of such knowledge-producing places is increasingly global, with prominent new nodes in China. Despite the diffusion of knowledge – and indicative of the peculiar nature of patents as an indicator of innovation – patents tend to be the most unequally distributed dimension of knowledge creation at the global level.

The dispersion of R&D has been a response to the location of both markets and talent, both of which have improved throughout much of the world as economic growth has taken place, particularly in Asia. As scientific and technological talent has improved in many places, the result on the ground is a range of capabilities, typically measured at the national level. Fagerberg et al. (2010) provide the most comprehensive review of how capabilities have been measured, incorporating one or more of several dimensions: science, research, and innovation; openness; production quality/standards; information and communication technology (ICT) infrastructure; finance; skills; quality of governance; and social values.

20.6.3 Keeping Track

A number of distinct efforts have been made to measure the technological capabilities of national economies, some for more academic interest and others for policymakers. Fagerberg et al. (2010) and Archibugi et al. (2009) compare many of these. Policymakers like such scoreboards for three reasons. First, they provide an “early warning system” for potential problems at a national level. Second, when used over time, national strengths and weaknesses can be monitored. Third, they help to focus firms, institutions, and government bodies on the same issues (Arundel and Hollanders 2008). Fagerberg et al. (2010) distinguish between several types of capabilities that indicate in various ways the capacity of the firms of a country to compete through creation of new technologies and to exploit existing knowledge from elsewhere.

It has become common for benchmarking and scoreboard reports to track the technological progress of national (and sometimes regional) economies. Indeed, there are so many scoreboards and sets of cross-national indicators that the EU produced its *Global Innovation Scoreboard* only in 2006 and 2008, now being content to publish only its *Innovation Union Competitiveness Report* and *Innovation Union Scoreboard* on an annual basis. All these measures are highly correlated – with one another and with gross domestic product (GDP) per capita (Fagerberg et al. 2010).

All of these address real – or imminent – technology gaps envisaged by the rise of China and other Asian competitors. All indicators and rankings are imperfect. Archibugi et al. (2009: 929) conclude, however, that “R&D intensity is less capable of explaining differences in innovative performance because *non-R&D factors* play an important role in differentiating national paths of innovation and performances”. However, there is a real risk that policymakers will drown in the flood of numbers from so many, especially annual, scoreboards.

20.7 Policy: Changing and Reacting to Changes in the Geography of Innovation

Here, policy refers to efforts at the national, regional, and local level to respond to – and to shape – the geography of innovation. Innovation policy has evolved from R&D alone to systemic – appreciating innovation as a systemic process. The OECD and, later, the EU have attempted to gather knowledge on the state of the art in policy and its empirical evaluation (Mytelka and Smith 2002). Policy continues to run ahead of theory (Steinmueller 2010).

20.7.1 Regional Innovation Policy: Constructing Advantage

Martin et al. (2011:566) provide convincing evidence that regional strategies based on one “best practice” model do not meet the very industry-specific needs of firms. In fact, these best-practice models . . . seem to be most well suited to industries that draw primarily on an analytical knowledge base”. Such sector-specific needs remind us of the importance of the nonspatial sectoral innovation systems. Any useful policy must include gatekeepers and other actors within a regional system who can interpret across sectoral and technological boundaries. These interactions work best when they are informal, untraded interdependencies rather than formal, contractual links. It is plainly difficult to create policy structures that must be at the same time formal (enacted in laws, personnel hired and evaluated, accounted for to taxpaying citizens) and informal (flexible and adaptable to new circumstances and knowledge).

Regions compete and, more than in the past, they work to create advantage in a world where the ability to attract and keep capital and people requires attention to infrastructure, institutions, policies, and innumerable details (Asheim et al. 2011; Cooke in Cooke et al. 2011). At a minimum, regional advantage should be

constructed more on the basis of the unique capabilities of firms and regions and not primarily on the basis of corporate or regional R&D efforts. What is needed is “smart specialization” (Lagendijk in Cooke et al. 2011). At worst, regions that fail to be attentive to these demands can become, or be perceived as, systemically innovation averse.

Constructed regional advantage, the current state-of-the-art regional innovation policy approach, takes into account three lessons from policy experience. First, *platform policies* represent “tailor-made policy strategies geared towards specific potentials and focused on tackling specific bottlenecks in regions that occur over time. As a result, regional policy needs to evolve, capitalizing on region-specific assets, rather than selecting from a portfolio of policy recipes that owed their success in different environments” (Asheim et al. 2011: 900; Cooke in Cooke et al. 2011; Harmaakorpi et al. in Cooke et al. 2011). Second, such a strategy must be based on *related variety*, rather than specialization or broad-based differentiation, to reflect shared and complementary knowledge bases and competences. The third element of this policy approach reflects that knowledge is distributed across traditionally defined sectors in distributed knowledge networks, and these knowledge bases are distinct and often incompatible with one another. Tura et al. (2008) illustrate several dimensions (structural, social, cultural, and intellectual) of *innovation platforms*, which reflect network-based innovative capability.

Policymakers, like firms, also face massive information overload. It appears that their ability to compete is made more difficult unless they use gatekeepers, such as consultants and service intermediaries who can help gather and synthesize knowledge from elsewhere. As with firms, the number of knowledge inputs to policy and the number of sources are increasing, and demands for data and synthesis – for example, for benchmarking – are common. Regional and national innovation policies now typically include university R&D, technology transfer, entrepreneurship, and spinoffs.

Regional absorptive capacity must be built and maintained, and it includes the absorptive capacities of firms located in the region, institutional features that promote knowledge exchange and learning in the region, and links to organizations elsewhere (Abreu in Cooke et al. 2011).

Vale (2011) dissects the standard policies related to clusters, which often downplay informal and untraded interaction among firms in an agglomeration. Further, he emphasizes that spatial localized learning processes are necessary but not sufficient for a successful cluster in a world where relevant knowledge is located in several – known and perhaps unknown – nonlocal and perhaps distant locations.

The complexity of innovation, not surprisingly, leads to complex frameworks for regional policy. Innovation policy generally is seen as messy and complex, with multiple levels and multiple actors including, in the European Union (EU), supranational policy. In this sense, innovation systems – like clusters – may be too difficult for policymakers to grasp fully and to coordinate adequately. Numerous intermediaries are involved at several levels (Nauwelaers 2011). As policy continues to run ahead of theory, specific programs are evaluated, but it is uncommon for technology policy to undergo evaluation (Steinmueller 2010).

Recent proposals for policy focus on the *cognitive* dimension of territories (Camagni 2009, Capello in Cooke et al. 2011), whereby territories act as learning regions (Simmie in Cooke et al. 2011). Knowledge-oriented policies (KOP) for regions as well as for firms can help to build competencies and to participate in the codevelopment of knowledge at a global scale. This involves several types of networks (Lawton Smith 2008).

Uyarra and Flanagan (2010) believe that the regional innovation system has become a fuzzy concept – attractive to policymakers and a useful “boundary object” linking but at the same time preserving the integrity of academic and policy discourses. The use of the term “system” encourages a view of regional economies as more-or-less closed systems and allows for inclusion of emergent, functioning, and dysfunctional systems. It also focuses attention on structure at the expense of agency.

20.7.2 Policy in a World of Global Production Networks and Global Value Chains

Manufacturing, long derided as a blue-collar sector staffed by uncreative people, of course includes engineers and other innovative personnel. The now-distant capabilities related to manufacturing leave many firms as “head-and-tail” companies with no body – the only activities remaining in-house are research and branding. Dankbaar (2007: 272) asks two pertinent questions: “Is there any reason to assume that research can be maintained as an in-house activity in the long run, if development and manufacturing have been outsourced? What happens to research if knowledge and experience coming from manufacturing and development are no longer immediately available?” TNCs make location decisions with a short-term perspective, but ultimately weaken the knowledge base of their home economies as suppliers of advanced materials, tools, production equipment, and components – collective capabilities – are no longer utilized as they also move or are replaced abroad.

Within global production networks, there has been a “geographic dispersion of cross-functional, knowledge-intensive support services that are intrinsically linked to production”. As flagship firms have moved to global sourcing, an “erosion of the collective knowledge which used to be a characteristic feature of the flagship’s home location . . . may have migrated for good to the supplier’s overseas cluster(s)” (Ernst 2002: 51, emphasis in original). In response to the new global situation, current advice for innovation policy is to frame such policy as a knowledge-based economy strategy, within a complex framework that includes the whole of government.

20.8 Conclusions

This brief survey has emphasized the systemic, learning-based model of innovation favored by many geographers and evolutionary economists. This view of innovation is able to embrace what we know about how innovation actually works – as a messy

and highly varied process that defies model builders. The standard model has not provided adequate guidance for policy, and this is why policy runs ahead of theory: it must do so but the result is that we have little systematic knowledge about how and why policies actually work. Policymaking often takes its cues from politics and political pressures rather than from empirical knowledge. Changes in the geography of innovation at the global scale affect regions and localities, both through the changing location of R&D and flows among nodes in the global system of knowledge. However, global forces are much more difficult for regional – and even national – policies to influence, as they are the outcome of independent choices by TNCs and by national policymakers. Innovative capability also is more difficult for any actor to assemble as technology grows more complex, and the necessary knowledge is found in ever more places.

Acknowledgment Thanks to Arnoud Lagendijk for his comments on an earlier version of this chapter.

References

- Acs ZJ, Braunerhjelm P, Audretsch DB, Carlsson B (2009) The knowledge spillover theory of entrepreneurship. *Small Bus Econ* 32(1):15–30
- Archibugi D, Denni M, Filippetti A (2009) The technological capabilities of nations: the state of the art of synthetic indicators. *Technol Forecast Social Change* 76(7):917–931
- Arundel A, Hollanders H (2008) Innovation scoreboards: indicators and policy use. In: Nauwelaers C, Wintjes R (eds) *Innovation policy in Europe: measurement and strategy*. Edward Elgar, Cheltenham, pp 29–52
- Asheim B, Gertler MS (2005) The geography of innovation: regional innovation systems. In: Fagerberg J, Mowery DC, Nelson RR (eds) *The Oxford handbook of innovation*. Oxford University Press, Oxford, pp 291–317
- Asheim BT, Boschma R, Cooke P (2011) Constructing regional advantage: platform policies based on related variety and differentiated knowledge bases. *Reg Stud* 45(7):893–904
- Audretsch DB, Keilbach M (2007) The theory of knowledge spillover entrepreneurship. *J Manag Stud* 44(7):1242–1254
- Camagni R (2009) Territorial capital and regional development. In: Capello R, Nijkamp P (eds) *Handbook of regional growth and development theories*. Edward Elgar, Cheltenham, pp 118–132
- Caraça J, Lundvall B-Å, Mendonça S (2009) The changing role of science in the innovation process: from queen to Cinderella? *Technol Forecast Social Change* 76:861–867
- Cohen WM, Levinthal DA (1989) Innovation and learning: the two faces of R&D. *Econ J* 99:569–596
- Cooke P (2005) Regionally asymmetric knowledge capabilities and open innovation: exploring ‘globalisation 2’ – a new model of industry organization. *Res Policy* 34:1128–1149
- Cooke P, Asheim B, Boschma R, Martin R, Schwarz D, Tödtling F (eds) (2011) *Handbook of regional innovation and growth*. Edward Elgar, Cheltenham
- Crevoisier O, Jeannerat H (2009) Territorial knowledge dynamics: from the proximity paradigm to multi-location milieus. *Eur Plan Stud* 17:1223–1241
- Dankbaar B (2007) Global sourcing and innovation: the consequences of losing both organizational and geographical proximity. *Eur Plan Stud* 15(2):271–288
- Ernst D (2002) Global production networks and the changing geography of innovation systems: implications for developing countries. *Econ Innovat New Technol* 11(6):497–523

- Fagerberg J, Srholec M, Verspagen B (2010) Innovation and economic development. In: Hall BH, Rosenberg N (eds) *Handbook of the economics of innovation*, vol 2. Elsevier, Amsterdam, pp 833–872
- Feldman MP, Kogler DF (2010) Stylized facts in the geography of innovation. In: Hall BH, Rosenberg N (eds) *Handbook of the economics of innovation*, vol 1. Elsevier, Amsterdam, pp 381–410
- Fischer MM, Revilla Diez J, Snickars F (2001) *Metropolitan innovation systems: theory and evidence from three metropolitan regions in Europe*. Springer, Berlin
- Freeman RB (2010) Globalization of scientific and engineering talent: international mobility of students, workers, and ideas and the world economy. *Econ Innovat New Technol* 19:393–406
- Griliches Z (1990) Patent statistics as economic indicators: a survey. *J Econ Lit* 28(4):1661–1707
- Lawton Smith H (2008) Inter-firm networks in high-tech clusters. In: Karlsson C (ed) *Handbook of research on innovation and clusters*. Edward Elgar, Cheltenham, pp 107–123
- Martin R, Moodysson J, Zukauskaite E (2011) Regional innovation policy beyond ‘best practice’: lessons from Sweden. *J Knowl Econ* 2(4):550–568
- Mason C (2008) Entrepreneurial dynamics and the origin and growth of high-tech clusters. In: Karlsson C (ed) *Handbook of research on innovation and clusters*. Edward Elgar, Cheltenham, pp 33–53
- Moodysson J, Jonsson O (2007) Knowledge collaboration and proximity: the spatial organization of biotech innovation projects. *Eur Urban Reg Stud* 14(2):115–131
- Mytelka LK, Smith K (2002) Policy learning and innovation theory: an interactive and co-evolving process. *Res Policy* 31(8–9):1467–1479
- Nagaoka S, Motohashi K, Goto A (2010) Patent statistics as an innovation indicator. In: Hall BH, Rosenberg N (eds) *Handbook of the economics of innovation*, vol 2. Elsevier, Amsterdam, pp 1083–1127
- Nauwelaers C (2011) Intermediaries in regional innovation systems: role and challenges for policy. In Cooke P, Asheim BT, Boschma R, Martin R, Schwartz D, Tödtling F (eds) *Handbook of Regional Innovation and Growth*. Edward Elgar, Cheltenham, pp 467–481
- Soete L, Verspagen B, Tel Weel B (2010) Systems of innovation. In: Hall BH, Rosenberg N (eds) *Handbook of the economics of innovation*, vol 2. Elsevier, Amsterdam, pp 1159–1180
- Steinmueller WE (2010) Economics of technology policy. In: Hall BH, Rosenberg N (eds) *Handbook of the economics of innovation*, vol 2. Elsevier, Amsterdam, pp 1181–1218
- Stoneman P (2010) Soft innovation: economics, product aesthetics, and the creative industries. Oxford University Press, Oxford
- Tura T, Harmaakorpi V, Pekkola S (2008) Breaking inside the black box: towards a dynamic evaluation framework for regional innovative capability. *Sci Public Policy* 35(10):733–744
- Uyarra E, Flanagan K (2010) From regional systems of innovation to regions as innovation policy spaces. *Environ Plan C Govern Policy* 28(4):681–695
- Vale M (2011) Innovation networks and local and regional development policy. In: Pike A, Rodriguez-Pose A, Tomaney J (eds) *Handbook of local and regional development*. Routledge, London, pp 413–424

Generation and Diffusion of Innovation

21

Börje Johansson

Contents

21.1	Introduction	392
21.2	Innovations and Heterogeneity of Firms and Places	393
21.3	Innovations, Diffusion, and Technological Development	396
21.3.1	Renewal Activities and Firm Performance	397
21.3.2	Characteristics and Performance of Firms	398
21.3.3	Proximity and Networks	400
21.4	Innovation, Regional Milieu, and Networks	402
21.4.1	A Functional Region Is an Arena for Face-to-Face Contacts	402
21.4.2	Urbanization and Localization	403
21.4.3	Accessibility to Knowledge Sources	404
21.5	Diffusion of Ideas and Technical Solutions	406
21.5.1	The Diffusion Model	406
21.5.2	Technology Diffusion and R&D Spillovers	407
21.5.3	Innovation Ideas and New Products	409
21.6	Conclusions	410
	References	411

Abstract

Generation and diffusion of innovation are two distinct processes that are interlinked in several ways. First, innovation efforts of firms are stimulated by the diffusion of innovation ideas. Second, the market penetration of successful product innovations diffuse to user firms and consumers, providing users opportunities to adopt novel routines and to imitate new designs. Third, creative destruction develops when a novel product finds its way to customers and replaces earlier product vintages, and this phenomenon has the nature of

B. Johansson

Department of Economics, Jönköping International Business School (JIBS), Jönköping, Sweden
e-mail: jobo@jibs.hj.se

a substitution process. All these processes are supported by knowledge flows which vary in intensity and diversity across the innovation milieu of functional regions. It is concluded that the milieu characteristics which stimulate innovation also stimulate adoption of novelties.

21.1 Introduction

The economic growth literature has, since the contribution of Solow in the 1950s, attributed productivity growth to processes of technical change, where economy-wide technical change is based on the generation of innovations and their diffusion across firms and regions, although innovation processes were not modeled as endogenous until the late 1980s (e.g., Romer 1986; Aghion and Howitt 1992). As firms develop new routines for their operation and design new products for the market, they do so with the objective to increase their productivity. Following suggestions in Nelson and Winter (1982) – building on Schumpeter (1934) – we shall consider three broad categories of innovations: (i) new *product varieties* with novel combinations of product attributes, (ii) new *firm routines*, comprising novel production and administration processes and techniques, and (iii) new markets including novel links to customers.

The research approach in the field of innovation studies has changed in important ways during the past two decades, primarily as a result of new data sources which contain firm-level micro data sets, allowing researchers to observe for individual firms' (i) firm characteristics, (ii) innovation efforts, and (iii) location characteristics. With information about the location of each firm, it becomes possible to consider information about the innovation milieu associated with different locations. An example of new data sources is the *community innovation surveys (CIS)*, in which data from the EU member states are collected on a regular basis with harmonized information (OECD 2005).

The objective of this chapter is to develop and examine a view on how firms generate innovations and what consequences innovations can have on firm performance and heterogeneity. A second task is to describe how innovations diffuse from innovators to other (user) firms and thereby affect the performance of the latter as well as the entire economy. In this endeavor the chapter presents a theoretical framework in which lasting differences in firm performance are related to persisting differences in firms' innovation and adoption behavior.

In the subsequent exposé, it will be shown that *regional milieus* that favor the generation of innovations also facilitate diffusion of novelties. As an example we may observe that a firm's generation of innovations is positively related to the knowledge intensity of the firm's employees. Likewise, the knowledge intensity of a firm's labor force increases its absorption capacity, augmenting the probability that novel techniques and product-attribute information will diffuse to the firm. The regional aspect follows when we observe that knowledge-intensive firms are more frequent in regions with a knowledge-intensive labor force.

A firm generates innovations in a process of innovation activities, of which *research and development (R&D)* efforts may be a major part. Innovation activities bring about new knowledge while at the same time using inputs from the conjunction of internal and external knowledge sources. The combination of internal and external knowledge accession is cumulated into knowledge about (i) firm routines, (ii) product variety attributes, (iii) markets and customers' willingness to pay for product attributes, and (iv) routines for how to organize and perform innovation activities.

The presentation will concentrate on three major aspects, namely, firm characteristics, innovation milieu characteristics, and innovation activities. The additional question is as follows: how do the three aspects affect firm performance, where these effects can be subdivided into direct and indirect innovation consequences? Direct effects concern new patents, markets and products, and sales of new products. Examples of indirect effects are factor productivity growth, sales per employee, labor productivity, and profitability. The reader should also recognize that the output from a firm is products, comprising both services and goods.

As a way to amalgamate established findings in the literature, the presentation will focus on the following set of theses associated with making innovations and adopting innovation:

Thesis 1: Innovation and adoption activities are two interlinked and overlapping firm renewal phenomena. As a rule both activities require that the pertinent firm renews its routines. In all essence, such routine adjustments should be classified as routine or process innovations.

Thesis 2: Both innovation and adoption activities are fuelled by inputs from internal and external knowledge sources. Part of the external knowledge is disseminated according to classical diffusion.

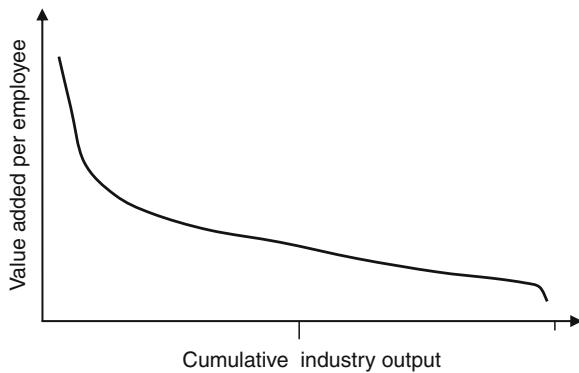
Thesis 3: Knowledge flows reduce in volume and intensity as the distance between origin and destination grows. This form of spatial discounting implies that localized knowledge is a fundamental characteristic of a functional region. In the sequel each functional region will be categorized with regard to its accessibility to different knowledge sources.

21.2 Innovations and Heterogeneity of Firms and Places

In order to maintain its competitiveness, a firm has to renew itself over time. As emphasized in Thesis 1, such renewal has two forms. The first is more proactive and is based on R&D and innovation efforts, where the firm in the spirit of Schumpeter generates product and process (routine) innovations. The second form of firm renewal is rather reactive and consists of a firm's search for novel product ideas and for new technical solutions developed elsewhere in the economy.

Continued and repeated *firm renewal* is the key factor for a firm's survival and eventual growth. In renewal efforts firms develop and adopt new technologies, using knowledge components that are paid for as well as knowledge that pass by as unintended consequences of current economic activities. To facilitate the combination of internal efforts and external interaction, firms can establish *networks*

Fig. 21.1 Labor productivity cross tabulated against output of firms in an industry (OECD code 1) in productivity-descending order



for knowledge flows. The associated spillovers and organized knowledge flows stimulate innovation as well as adoption activities of firms and provide inputs to an ongoing process of firm renewal. Differences in firms' renewal intensity will cause heterogeneity among firms to sustain. Thesis 3 clarifies that proximity to knowledge sources matters.

Firms in the same industry or firms supplying product varieties belonging to the same product group have as a rule heterogeneous characteristics and display different performance in terms of productivity, profitability, or growth, and they differ in their R&D and innovation efforts. In established microeconomic theory, such differences are predicted to vanish over time, based on the argument that only the best practice can survive. Empirical observations do not support this view (Dosi and Nelson 2010). To a large extent, interfirm differences remain over long time sequences.

Differences between firms in a given industry (or group of industries) may be identified for a panel of firm observations over time. Such a panel will contain differences for each individual firm at different points in time as well as differences between firms that remain basically unchanged along time. With reference to Geroski (1998), we may then calculate the total variance for a performance variable like value added or gross profit per employee and then in a second step determine how much of the variance is due to the variation between years for each individual firm, referred to as *within variance*. The remaining variance can then be conceived as a persistent difference between firms, referred to as *between variance*. This between variance is typically 3–4 times larger than the within variance (Andersson et al. 2012). Such observations demonstrate heterogeneity among firms in most industries while at the same time showing that differences between firms persist over time.

Figure 21.1 provides a picture of the labor productivity (2006) in an industry supplying differentiated products (OECD code 1), based on Swedish data. The horizontal axis measures the cumulative output from firms in the sector when firms are ordered according to descending productivity. The figure illustrates how the quartile with the highest productivity has a productivity which is 3–4 times as large as the lowest quartile. Such performance differences provide a strong motive to examine firm characteristics when assessing performance. Among such characteristics the literature has considered firms' behavior with regard to efforts to innovate and

to adopt new technology developed by other firms. The first aspect associates with the process of generating innovation and the second with diffusion and adoption.

The presentation in this chapter makes use of a theoretical framework, in which lasting differences in firm performance are related to persistent differences in how firms generate their own innovations and how they adopt new equipment as well as current input flows. Thus, in a second perspective, we observe that there are also systematic differences between firms with regard to how large amounts of resources per sales and value added they commit to innovation efforts (R&D intensity). Such R&D shares display highly skewed distributions across firms, and the firm differences remain persistent over time (Klette and Kortum 2004). The picture that emerges is a system with different “species,” where a large share of firms is not engaged in innovation and R&D activities, where some firms are innovation active only occasionally, whereas firms in still another group remain persistently innovation active over a sequence of years. In addition, firms that display alertness in buying or imitating innovations developed elsewhere can be expected to have a higher *absorption capacity* than the average firm.

As firms are located in different regions, one may also investigate to what extent a firm’s local economic milieu affects its performance. Are there characteristics of a firm’s economic environment that influence firm performance both in terms of economic outcome and innovation results? Before that question can be answered properly, it is necessary to consider how a region is defined and identified. This presentation refers to the concept *functional urban region*, where the regional boundaries encircle an area within which frequent face-to-face interaction can take place with short notice and without travel planning. In most practical cases, this requirement is satisfied for local labor market regions, for which labor market commuting between intra-regional areas is much more intensive than between areas in two different functional regions.

Heterogeneity among firms is a result of different development paths, and these paths are consequences of how well each firm manages to carry out its own innovations and to adopt technological novelties developed elsewhere in the economy. Innovation efforts of firms combine with knowledge flows of various kind including interactive communication between the firm and other actors such as suppliers, customers, competitors, university researchers, and other knowledge providers. Knowledge flows are equally important for a firm’s efforts to acquire and adopt innovations. In the case of adoption, the driving incentive is to learn about new equipment and technical solutions as an input to making adoption decisions, and this makes knowledge flows vital. Similar observations apply to firms which imitate novel products with new attribute combinations.

Knowledge flows vary in content, diversity, and intensity between functional regions. In particular, the friction of knowledge diffusion and transfer is smaller inside a functional region than more long-distance flows. In this way we can explain the pronounced tendency of innovations and technology adoption to cluster in particular functional regions, caused by the heterogeneity of urban regions with regard to knowledge intensity of the labor force, the presence of R&D activities in firms and universities, the size of gross regional product (GRP) and level of GRP

Table 21.1 Illustration of urban region heterogeneity in Europe 2004

Rank order (GDP/cap)	Urban region	GDP/cap €	GDP €, million	Accessibility to GDP
1	Paris	67,500	146,000	823
2	Inner London	65,600	191,300	815
11	Copenhagen	38,300	46,100	178
33	Stuttgart	30,400	121,300	315
87	Glasgow	26,000	49,800	81
167	Skåne (Scania)	23,700	27,500	76

Remark: 640 NUTS 2/3 regions with PPS-adjusted GDP values (Eurostat)

per capita, the diversity of the region's export and import flows, etc. For the USA, such differences between urban regions are recognized and documented in, for example, Henderson (1997).

In their evolution urban or city regions remain different from each other and retain their idiosyncratic features including markedly different levels of income per inhabitant. As illustrated in Table 21.1, urban regions also differ in their human capital resources, which is reflected in the table by the knowledge intensity of the labor force, measured as the share of the labor force with at least 3 years' university studies. Three things are accentuated in the table. First, the *knowledge intensity* increases very fast between 1993 and 2007. Second, the relative difference between regions remains unchanged during the 14 years. Third, the larger the urban region is, the higher its knowledge intensity. Subsequently the presentation will emphasize a region's knowledge intensity as a determinant of its innovation as well as adoption intensity. High knowledge intensity is associated with two region features: (i) high absorption capacity of firms and (ii) intense knowledge flows and spillover phenomena, where pure spillovers are defined as unintended knowledge flows which occur free of charge.

21.3 Innovations, Diffusion, and Technological Development

As summarized by Mansfield (1987), technological change results in a change in the production function of an existing product (routine renewal) or in an addition to the list of technologically feasible products (product renewal). In practice both types of renewal may occur simultaneously. A firm's adoption of novelties may not be very different. As a firm purchases new equipment or imitates product innovations made by others, it may have to redesign its routines.

Figure 21.2 provides a stylized picture of how a firm renewal can affect firm performance and how a firm's innovation and adoption activities depend on:

- The characteristics of the firm which includes its innovation capabilities and innovation strategy
- The characteristics of the firm's *innovation milieu*, where the latter primarily corresponds to the possibility of knowledge interaction in the functional region where the firm is located

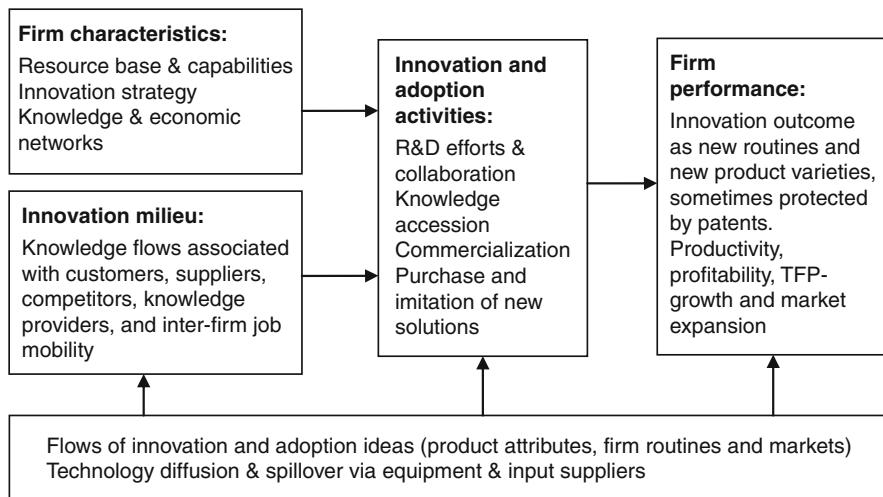


Fig. 21.2 Factors influencing the size and consequences of innovation and adoption activities

The figure illustrates the statements in the first and second theses of this chapter and attempts to connect the basic concepts for the analysis of innovation and adoption activities while putting the variety of diffusion processes in focus. First, as observed in the figure, diffusion of innovation and adoption ideas stimulates firms to make innovation efforts and to adopt new technical equipment. In addition, when a firm introduces a product innovation, then the market penetration of its new product can be modeled and analyzed as a diffusion process. Diffusion of new products that are sold to firms and households or imitated by user firms also represents economy-wide renewal which is usually labeled technological development. It also associates with *product cycle* dynamics (Vernon 1966).

21.3.1 Renewal Activities and Firm Performance

Figure 21.2 depicts how innovation activities can influence *firm performance*, subdivided into innovation outcomes and economic outcome. The first type of performance focuses on the innovation output of a firm, and the associated performance indicators comprise measures such as number of new products and patents, sales of new products, and sales of new export products. These are direct outcomes of renewal efforts to innovate and adopt novelties in the economic environment. The second type of performance refers to indirect consequences of renewal activities, measured by the economic result of the firm as reflected by the level and change in productivity and profitability, growth of total factor productivity (TFP growth), and increasing value of the firm and its market expansion. These are all different returns to renewal activities. As will be evident in Sect. 21.5 of this chapter, the economy-wide returns to an industry's R&D are greater than the

returns to the typical single firm. As recognized early by Mansfield (1968), this is due to the role played by technology diffusion and spillovers between firms and industries as firms buy improved equipment and intermediaries.

An important message in this presentation is that a firm's development is the result of its *renewal activities* which include innovation and adoption. These two phenomena overlap without a clear demarcation line. We first observe that a firm that introduces new product varieties also has to adjust its production processes and thus make a routine innovation, indicating that product and process innovations are complementary (Nyström 2006). Such routine innovations are equally probable, irrespective of whether it is a matter of a product innovation or an imitation. Moreover, the adoption of new technical solutions in the form of new equipment also stimulates the firm to make routine innovations. Hence, innovation and adoption are intermingled. This reflects the chapter's second thesis saying that innovation and adoption activities are part of a firm's overall renewal efforts, and this view is reflected in the CIS specification of innovation efforts, which includes the following list of innovation and adoption activities: (i) internal R&D work, (ii) external R&D work, (iii) acquisition of new equipment and associated cost, (iv) accession of external knowledge, (v) allocation of time for human resource development and training, (vi) marketing and commercialization efforts associated with new products, (vii) product development and design, (viii) development and maintenance of links to external actors for R&D collaboration, and (ix) scanning external knowledge sources for innovation and adoption efforts (OECD 2005).

A firm's possibilities to carry out renewal activities depend in a critical way on the resource base and associated capabilities of the firm. These are durable and difficult-to-imitate capacities. The literature in this area (e.g., Teece 2010) emphasizes the capability of an innovating firm to develop, maintain, and orchestrate its resource base to adapt in an ever-changing business environment. The *knowledge intensity* and the firm's experiences from previous innovation and adoption efforts are the core determinants of firm renewal. The associated knowledge assets are based on learning how to organize and establish *routines* for conducting innovation activities.

21.3.2 Characteristics and Performance of Firms

Firm characteristics can be organized under the headings *strategy*, *renewal capabilities*, and *networks*. The characteristics influence what the firm is capable of doing but also what it intends to do. Intentions and objectives may be reflected by the firm's innovation strategy which comprises the firm's commitment to systematic R&D and its ambitions to develop capabilities and networks for knowledge flows over time. Recent studies suggest that firms display permanent heterogeneity that can be grouped into no, occasional, and persistent engagement in R&D and other innovation activities. With persistent engagement, the firm is rewarded with learning routines for how to conduct R&D, and this leads to firm knowledge and experiences that improve performance. Another strategy aspect is the size of innovation expenditures, often proxied by R&D intensity.

R&D intensity is usually calculated as the ratio between R&D expenditures and sales or between R&D expenditures and value added. In studies of the impact of R&D on firm performance, the measure R&D intensity controls for size. With this approach heterogeneity is revealed by the observation that as much as half of the variation in R&D intensity is explained by fixed firm effects. In view of this, one may remark that the two most widely used indicators of firm characteristics – cash flow and the degree of diversification – explain much less of interfirm differences in R&D intensity. Thus, the fixed firm effects indicate strongly that differences between firms' innovation efforts have a tendency to remain invariant over time. This can also be interpreted as a finding saying that firms employ different innovation strategies.

Having reached this point, one may ask: what about Schumpeter's suggestion that firm size affects its innovation intensity? Cohen and Levin (1989) and many other survey contributions suggest that it is difficult to reject the hypothesis that R&D efforts are proportional to firm size. Andersson and Johansson (2010) argue in a model with product variety innovations and multiple export markets that firms are large as a consequence of variety and market innovations in the past, not the other way around. Instead, cumulated innovation experiences affect a firm's probability of innovating in the future.

In order to develop its *innovation capabilities*, a firm has to invest in such capacities, and as a consequence we find that firms with a large share of knowledge-intensive employees spend more than average resources on innovation activities, and they do this in a more persistent way than the average firm. A firm with an offensive and sustainable strategy of this kind may also have more favorable economic outcome than average. Obviously, this also reflects dynamic interdependencies with ambiguous causation. In cross section as well as panel data analyses, one can observe that the likelihood of making innovation efforts is associated with the same variables as those which are associated with the economic result of these efforts. Such variables which correlate with higher innovation intensity and higher returns to innovation include (Andersson et al. 2012):

- Knowledge intensity of the labor force (human capital).
- Physical capital.
- Repeated innovation efforts.
- The firm belongs to a multinational company group.
- Market extension and export experiences.
- Import intensity and import links to foreign suppliers.

Instead of extending the review to consider other ways of relating firm performance to characteristics of firms, the presentation will focus on the conditions enumerated above to examine their association with location characteristics and innovation milieu of the innovating firm. Empirical observations suggest that the intensity and composition of knowledge flows are basic in explaining a firm's innovation engagement and economic return to its efforts. Table 21.2 provides an overview of different knowledge sources, of mechanisms influencing generation and transfer of knowledge of various character, and of spatial aspects of these mechanisms.

The table traces a broad set of *knowledge sources*, including knowledge exchange with collaborators' purchase from knowledge providers, pure knowledge spillovers as a side effect of ordinary transactions, knowledge that moves from one firm to another

Table 21.2 Knowledge intensity in the private sector of the economy. Sweden 2007

Functional region	The entire private sector 1993, %	The entire private sector 2007, %	Inhabitants 2007 (region average)
Stockholm metropolitan region	17.3	28.1	2.3 million
Göteborg – Malmö regions (average)	13.6	23.8	1.0 million
Medium-sized urban regions	10.3	18.6	0.2 million
Country average	8.6	14.7	0.1 million

Remark: Knowledge intensity is the share of the labor force with at least 3 years of university studies.

Source: Elaborations from Statistics Sweden (Johansson et al. 2010)

when persons switch from one employer to another, entrepreneurs that leave employment and start new firms, active search for *knowledge-accession* possibilities, and knowledge flows in long-distance networks such as internal links inside a multinational corporation. Many of these phenomena can be understood as diffusion of creative ideas, and we can identify network externalities when firms establish network links to carry both intended and unintended knowledge flows.

21.3.3 Proximity and Networks

While relying on its cumulated resource base and associated knowledge assets, the innovating firm is characterized by its capacity to exploit in-house knowledge in conjunction with external knowledge sources. The latter are characterized in Table 21.3, from which it is obvious that a firm's proximity to external knowledge affects the opportunities to acquire useful inputs to the firm's innovation activities.

As outlined in Johansson and Quigley (2004), there are two principle ways that can simplify and stimulate knowledge interaction and exchange of associated information. The first principle is the *proximity advantage*, which is based on the fact that the frequency of face-to-face (FTF) interaction between two or several parties decreases as the (time) distance between the location of the parties increases. This principle implies that an innovating (and adopting) firm benefits from being located in an environment with rich and diverse knowledge flows and with a multiplicity of relevant knowledge sources and knowledge exchange actors like R&D-intensive firms, knowledge-intensive producer services, and research organizations, including universities.

The second approach to facilitate knowledge exchange between two parties is to invest in links (communication channels) between the parties. According to this principle, a firm can invest in links and entire networks of interaction links to reduce the friction and costs of interaction over long distance. This opportunity may be termed *network advantage*. Thus, when a proximity solution is not at hand in a given location, then a firm can choose to invest in links to distant collaborators (such as suppliers, customers, and other knowledge providers) as a means to compensate for

Table 21.3 Origin of knowledge flows that are inputs to a firm's innovation activities

Origin or source of knowledge flows	Generation and transfer of knowledge	Spatial aspects and co-location in the same region
Knowledge interaction	Collaboration with customers, suppliers, universities and other knowledge providers	The interactive efforts are facilitated when partners are co-located in the same region
Purchase of knowledge (e.g., from knowledge-intensive producer services suppliers)	Knowledge transactions may require links of trust between buyer and seller	Location proximity facilitates the establishment of contract-like links between actors
Spillovers from normal transactions between a firm and its customers and suppliers	The firm's interplay with customers, suppliers and other actors open up for unintended knowledge flows	A firm's transaction links extend across region and country borders, but intra-region links are more likely to establish in large urban regions
Job mobility bringing the firm new labor embodying knowledge achieved in previous job(s)	Recruitment inflow to a firm may be the basic source for unintended spillovers. Such flows decline with increasing distance	The frequency of job switching is more frequent (i) among knowledge-intensive labor and (ii) in large urban regions
Scanning and searching for knowledge accession opportunities	Renewal in the form of innovation and adoption is fuelled by the conjunction of internal and external knowledge	Firms located in urban regions which host many and diverse knowledge sources offer the local firms external knowledge advantages
Internal knowledge flows between units of a company group, especially multinationals	The internal networks of a multinational company group can overcome long distance and protect knowledge from leakage	The multinational subsidiaries can engage in knowledge accession and local networks in selected global set of nodes
Investment in R&D collaboration networks locally and globally	These networks include strategic alliances as well local links based on trust	Collaboration links reduce the friction of knowledge exchange and the payoff becomes higher and longer the planned interaction frequency is

the lack of feasible proximity options. In many cases lumpy investments in long-distance links complement investments in links for short-distance interaction. The advantage of a location in an agglomeration is that (i) the need for lumpy link investments is smaller in an urban agglomeration, while such investments at the same time are more easy to establish inside an agglomeration. In particular, when two actors are located in the same functional region, the cost of forming an interaction link should generically be smaller than when the same actors are more distant from each other. This conclusion can be motivated in the following way:

Consider that two actors strive to develop a mutual interaction link which has the form of an implicit contract underpinned by trust (based on positive experiences) and a joint capacity to communicate complex messages in a reliable way. We can assume that such relations require repeated face-to-face (FTF) contacts between the two parties as an input to the link formation, while maintenance

comes naturally as a consequence of using the link. In this view, a link is less costly to establish for firms which are colocated in the same region than for firms hosted in different regions.

A firm can be defined as *innovating* for periods during which it is engaged in innovation activities. Innovating firms have an advantage from being located in a large agglomeration with many opportunities to interact and many opportunities to establish local interaction links. This observation opens questions about the geography of diffusion, which is a field where a major contribution was made by Hägerstrand in the beginning of the 1950s. Hägerstrand (1967) used a huge set of observations to demonstrate the statement in the third thesis of this presentation. This statement stresses that ideas, production methods, and new products diffuse across geographical areas in spatial processes exhibiting clear regularities, where the novelties diffuse faster along short distances from the original source. In this sense the Hägerstrand model has been considered to stress neighborhood effects in the spreading process while observing that large and dense places represent a greater potential of neighborhoods.

Hägerstrand investigates several alternative explanations of innovation diffusion processes. For example, one may assume that (i) the entire population (of potential adopters) becomes informed about the innovation simultaneously, whereas acceptance of the novelty occurs in a random order of precedence. This may be varied by considering unevenly distributed capacities to accept the novelty, and this may in turn be associated with the presence of “innovation centers” and followers ordered in a hierarchy. From this we may conclude that (i) if receptiveness or propensity to adopt is unevenly distributed, spatial diffusion will unfold accordingly, and (ii) if the generation of novelties is more frequent in certain places, neighborhood effects will affect the *spatial diffusion* pattern.

21.4 Innovation, Regional Milieu, and Networks

Empirical observations suggest that innovation is spatially concentrated. Innovation combines invention and commercialization, and this may explain why innovation is more concentrated than invention and more concentrated than production. However, the basic observation in this section is that knowledge is spatially sticky. In every particular case of *knowledge diffusion* (spillover as well as commercial transfer), the friction cost will vary because of communication distances. This friction is augmented when knowledge is complex (Beckmann 2000) and when it is tacit (Polanyi 1966). In both cases messages are difficult to encode and decode, and the tool to overcome this obstacle is frequent *FTF interactions*. This makes knowledge spatially sticky (von Hippel 1994).

21.4.1 A Functional Region Is an Arena for Face-to-Face Contacts

In previous sections of this chapter, the presentation argues that an innovative firm has to rely on both internal knowledge workers and the presence of knowledge-intensive labor in the environment. A firm’s accessibility to knowledge intensity in its nearby

environment can benefit the firm in two different ways. First, a large local supply of labor with university education facilitates the matching of supply and demand with regard to qualifications and competence profiles. The second aspect is that a region with a wide spectrum of knowledge resources provides rich opportunities for knowledge exchange and creative interaction with other actors in the urban region.

The above observations refer to both pecuniary and other knowledge flow externalities. First, transaction costs for recruiting employees with a desired profile reduce in a large urban region. Second, transaction costs also reduce in processes of knowledge accession. Third, pure *knowledge spillover* can be expected to increase as the size of an urban agglomeration expands. A large urban region can afford diverse and frequent FTF contacts at low costs, and this explains the reduction of knowledge transaction costs and the augmented likelihood for spillovers.

Ohlin's early discussion of urbanization economies was reemphasized in the contributions by Jacobs (1969), where large urban agglomerations are depicted as places with diversity in competence, ideas, product innovations, and variation-rich import flows. Such milieus, Jacob argues, foster creativity and innovation activities, especially since they concretize Schumpeter's vision of *novelty by combination*.

If urbanization economies obtain in a milieu of complex diversity, localization economies may be characterized as a milieu with a spectrum of input suppliers and other support factors that are designed to improve colocated firms in the same industry, firms supplying varieties that belong to the same product group or firms which share the same categories of customers and suppliers. Cluster milieus with localization economies may be considered as the agglomeration phenomenon that can develop in small- and medium-sized urban regions, whereas urbanization economies is a characteristic of large urban (metropolitan) regions. In Capello (2002), it is argued that industry clusters are prevalent in large urban regions, while observing that in a metropolitan region there can be many types of clusters, making the economy a "cluster of clusters."

Especially for *cluster* phenomena, the literature has stressed the role of communication links between firms extended to complex networks for knowledge exchange among firms in the same cluster. A prerequisite is of course that the pertinent firms must have enough knowledge to exchange. In this view the network is rather an infrastructure for product and process development activities.

Firms belonging to a multinational company group have the internal network of the group as an infrastructure for knowledge interaction. First, such company group networks are especially designed to protect knowledge from leaking to competitors in undesired ways. Second, the global location of subsidiaries makes it possible for individual firms in a group to tap knowledge from different knowledge centers around the world.

21.4.2 Urbanization and Localization

Agglomeration of firms can theoretically be divided into two forms. The first case is obtained when several firms in the same industry collocate or cluster in the same urban region. In the second stance, agglomeration refers to colocation in the same

urban region of firms that belong to different types of industries. Clustering of similar firms is assumed to bring about localization economies with diversity within a specialized field, whereas agglomeration of firms adhering to a variety of industries is assumed to cause urbanization economies, where size and diversity of demand are expected to attract a diverse supply. This distinction between localization and urbanization economies was made by Ohlin (1933) and was later studied by Henderson (1997) and many others.

When a set of firms in the same industry are colocated in the same functional region, they benefit from *localization economies* due to mutual stimuli to improve production routines and to develop novel products. The consequence of such colocation can be augmented productivity of the pertinent firms. Localization economies are often thought of as an externality generated by colocation of several firms that have similarities with regard to markets (customers), intermediary inputs, technology and equipment, distribution systems, and the like. Having much in common, those colocated firms can mutually exchange and spill over adoption and innovation opportunities and technical knowledge. This phenomenon can be expected to have a significant role to play in smaller (urban) regions which may develop an environment of interlinked firms and their specialized suppliers of services and other inputs and associated institutions like trade associations and universities.

Obviously, a successful regional cluster may in the long term be affected by negative lock-in effects, such that they develop into mutual stiffness while evolving along a life-cycle path, starting with a juvenile period of expansion, followed by stagnation and eventually decline. In contradistinction we observe that large urban agglomerations in principle are protected against this phenomenon by having a broader spectrum of specialized fields and diversification as its basis. As emphasized by Jacobs (1969), the urban diversity constitutes an environment that boosts creativity and opens an avenue that facilitates the cross-fertilization of ideas. A very similar view was put forward by Vernon (1966) when he suggested that new product cycles frequently are initiated in metropolitan regions with rich knowledge sources, intense knowledge flows, and competent and demanding customers side by side with alert input suppliers. In this view innovations are generated where *urbanization economies* prevail and foster communication externalities.

A long range of empirical studies can be summarized by suggesting that large urban agglomerations are more innovative while at the same time being among the most productive places. These studies also suggest that metropolitan-region advantages are caused by economies of scope. These regions attract talented persons with creative occupations to migrate into metropolitan regions, and hence it becomes troublesome to which extent higher productivity and higher wages are caused by a metropolitan region's productive milieu or by a selective in-migration of skilled persons.

21.4.3 Accessibility to Knowledge Sources

Consider an economy which consists of a set of urban regions, $r \in R = \{1, \dots, \bar{r}\}$, as specified earlier in this chapter, and assume that each region consists of one or

several urban areas, $i \in r$, where each such area, however small, represents a spatial concentration of economic activity. Many early studies have examined how aggregate knowledge sources and R&D activities inside an urban region generate spillovers and affect innovation activities and innovation outcome of firms located in the region. The conclusion from many of these contributions is that knowledge flows and *spillovers* are spatially bounded in the sense that the likelihood of knowledge flows reduces as distance between origin and destination grows.

Let G_i be the amount of a knowledge resource located in urban area i , and consider that we can measure the time distance, t_{ij} between any two pairs of locations i and j . Let λ be a parameter reflecting the time sensitivity of FTF contacts between the two locations, and assume that $\exp\{-\lambda t_{ij}\}G_j$ is a measure of the potential for knowledge flows (including spillovers) from knowledge sources in urban area j to area i . This formulation, which can be derived from a random-choice specification (Andersson and Gråsjö 2009), implies that the potential for knowledge flows on the link (i, j) reduces in value as the time distance increases. The total knowledge flow potential of firms in urban area i , A_i , can be calculated as

$$A_i = \sum_r \sum_{j \in r} \exp\{-\lambda t_{ij}\}G_j \quad (21.1)$$

while the intra-regional potential equals $A_i^r = \sum_{j \in r} \exp\{-\lambda t_{ij}\}G_j$. One may interpret A_i as the overall *accessibility* to knowledge sources of firms in urban area i , while A_i^r represents the intra-regional accessibility to knowledge sources for firms in area i in region r . The measure of A_i in Eq. (21.1) represents an alternative to using an aggregate G-value for an entire urban or metropolitan region. In particular, the A_i -measure is not based on an arbitrary (administrative) delineation of the boundaries of an urban region.

Andersson and Gråsjö, (2009) employ a model with a knowledge production function (KPF), with patent applications of firms representing output, whereas internal and external knowledge sources comprise the inputs. The knowledge production is assumed to depend on R&D activities (man years) in other firms and R&D activities in universities (man years). The influence from these external knowledge resources is discounted according to the principle described in Eq. (21.1), but separated into local, intra-regional, and extra-regional influences. The study demonstrates that such an accessibility approach takes care of the spatial interdependencies by including them in the model. The described approach demonstrates a way to model spatial *knowledge interaction* opportunities, and for each given place, it provides a measure of the potential for diffusion of ideas from the surrounding environment to the selected place – in line with the model contributions of Hägerstrand as presented in Sect. 21.3.3. With another econometric technique, Fischer and Varga (2003) also provide evidence in favor of Hägerstrand's conclusion about distance decay effects.

21.5 Diffusion of Ideas and Technical Solutions

Technological diffusion has primarily been studied in the narrow perspective of firms starting to use a new method of production or, in other words, the adoption of a new production technique. Mansfield (1968) observes that until the end of the 1950s, economists allocated little attention to factors that may determine the rate of diffusion of a new technical solution. Studies during the following decades brought a considerable amount of information about the rate of diffusion (see Batten and Johansson 1989). All these studies also confirm that the diffusion approximates a sigmoid pattern, which can be mimicked by a logistic equation.

Technological diffusion of the kind referred to above plays a prominent role in a country's economic development, since it spreads new technical solutions that are applied across firms and regions. The diffusion can have the form of an innovation that is commercialized and sold in the form of new machinery equipment, information and communication technology (ICT) routines for logistics and administration, etc. In this case the diffusion is partly the result of marketing and sales efforts made by the innovator. The innovation may also originate from a firm which makes an invention to be used in the firm's own operation, while this innovation diffuses to other users through imitation. In this second case, the diffusion may take place in spite of efforts from the innovator to keep the innovation secret.

21.5.1 The Diffusion Model

Technology diffusion is just one of several diffusion processes that have been studied. We may, for example, consider the “epidemic” diffusion of social norms and consumer behavior. In the sequel we will consider diffusion of innovation ideas, representing knowledge that can be input to firms' efforts to develop new product varieties. Moreover, imitation of new products across firms and regions also has the character of a diffusion process. Such spread processes include the establishment of Chinese restaurants in cities over the globe in the 1960s and 1970s, as well as Sushi bars in the following decades.

Irrespective of the nature of what is being diffused, empirical observations provide evidence that the process follows a similar pattern in almost all cases. To make this obvious, a share variable, z , can be introduced, where for each point in time, z refers to (i) the share of a firm's operations that makes use of a specific technique, (ii) the share of all potential users of a new technique (or a new type of current input) who are employing the new technique, (iii) the share of a specific market that a new type of product has managed to conquer, or (iv) the market share a firm has obtained for a specified product segment. Frequently the variable z is referred to as the share of adopters or the market penetration share.

Consider now the development of z as described by the following differential equation:

$$\dot{z} = cz(1 - z) \quad (21.2)$$

where $c \geq 0$ is a given constant that describes the speed of change and $\dot{z} = dz/dt$. When z is small, the share of non-adopters, $(1 - z)$, is large, thereby providing opportunities for the novelty to “randomly” find potential adopters. As z grows and becomes larger, the uncertainty about the novelty is reduced and the propensity to adopt goes up, although the share of non-adopters, $(1 - z)$, is gradually becoming smaller. The outcome is the well-known sigmoid curve in Fig. 21.3, as depicted by graph A.

The Verhulst equation in Eq. (21.2) describes a logistic growth path of, and it can be solved to obtain $z/(1 - z) = \exp\{c(t - \tilde{t})\}$, where \tilde{t} denotes the time it takes for $z(t)$ to reach the value 0.5. Rearranging once more, the share at time t has the value $z(t) = [1 + \exp\{-c(t - \tilde{t})\}]^{-1}$.

For the model of technology *diffusion*, $z(t)$ can for an industry reflect the share of all potential adopters which at time t have already started to use the new technical solution. For the individual firm, z may instead measure the ratio between how much the firm has installed of the new technique relative to a complete installment.

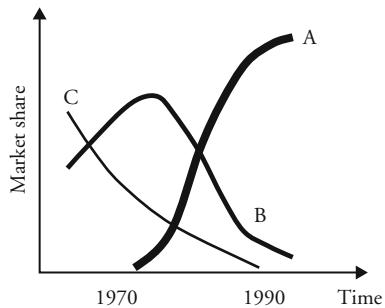
In a seminal study by Griliches (1957), the diffusion of hybrid corn during 1932–1956 is shown to follow *S-shaped* paths for each of five states in the so-called Corn Belt of the USA. Moreover, the introduction of this novel hybrid seed method took place in a sequential order, with Iowa as the initiator or forerunner and with other states following with different time lags vis-à-vis Iowa. The study provides clear indication that the introduction started earlier for states in which the new method had higher profitability, and it also spread at a faster pace in places with higher profitability.

21.5.2 Technology Diffusion and R&D Spillovers

A firm’s production is based on internal resources and on inputs (equipment and intermediaries) bought from other firms. Typically the intermediaries account for at least half of the sales value, where outsourcing strategies of firms leads to an increasing share of intermediary inflows and a reduced share of value added. By means of its own R&D, the individual firm can increase its value added and reduce its cost of intermediary inputs. This type of change process improves the firm’s performance over time. However, there is a parallel process which also affects firm performance. Firms that supply inputs also spend resources to improve their deliveries, and these improvements have the capacity to affect input-buying firms’ performance. This secondary effect has been labeled R&D spillovers.

R&D spillovers refer to the direct knowledge gains of input-buying firms from the R&D of input-supplying industries. An early contribution to this form of analysis is from the 1960s, followed by studies that calculated measures of the amount of R&D embodied in customer firms’ inputs, based on information about capital purchases made by one industry from other industries. A step further was taken in the 1980s in studies using the product R&D made by input suppliers to obtain a measure of R&D spillovers.

Fig. 21.3 Market share development for (C) mechanical typewrites, (B) electrical typewrites, and (A) word processing equipment. Sweden 1970–1990 (Johansson and Karlsson 1991, p 21)



Wolff (1997) applies a measure of *embodied* technical change which is a weighted average of the TFP growth of the supplying industries, using the customer industry's input-output coefficients as weights. This can be referred to as direct productivity spillover and reflects the idea that spillover flows are proportional to inter-sector flows. The same study indicates an even stronger effect on an industry's rate of technical progress when input flows are constrained to be equipment deliveries. Results of this kind seem to indicate that disembodied knowledge flows follow the same patterns as equipment-embodied technology diffusion.

Technology spillovers is a partly misleading notion since an important share of knowledge flows through the economy has the form of purchased knowledge, often embodied in equipment and systems that are acquired by and installed in user firms. Two aspects are important to contemplate. First, when new equipment and new types of current inputs are being developed, the purchasing firm has to find out which are the suppliers offering suitable solutions and which are not. In this context certain firms are more capable and have more advanced absorption capacity. This may be described as firms' search for input suppliers that offer the best practice. This type of knowledge search and accession is not R&D but has many characteristics in common with innovation efforts. In particular, when a firm assesses new equipment options, it may also have to consider new routines and technical solutions and add own innovation efforts.

Second, the opportunities to scan and collect information about input alternatives and novel equipment vary between each innovation milieu associated with a specific location, where innovation milieu signifies the localized knowledge-accession and innovation-collaboration opportunities in the environment of the location. In this context a location is identified as a particular urban region. In an environment of alert input suppliers, the likelihood of finding relevant input alternatives is generally higher than when information has to be collected from more distant sources. Proximity to suppliers brings greater opportunities to communicate and interact with established and potential suppliers. In this way *proximity* may lead to the formation of interpersonal communication networks that can facilitate learning and development. Adams and Jaffe (1996) found that the effects of parent R&D on plant-level productivity were diminished by geographic distance as well as with technological distance, providing further evidence in favor of Thesis 3 in this chapter.

In regional science the importance of the innovation milieu has been studied for a long period, with one strand focusing on cluster formation in smaller urban regions – also referred to as localization economies – and a second strand following the theory of agglomeration economies – also recognized as *urbanization economies*. The arguments for how cluster and agglomeration economies foster technology diffusion and adoption run parallel to those focusing on the generation of innovations. A *cluster* may be rich in specialized input suppliers in view of a particular industry. Agglomeration economies offer diversity of different categories of specialization and can, as a consequence, foster a richer variety of novel combinations.

As we have seen, the sources of technology diffusion and spillover can be fostered by a milieu of local suppliers that have clustered in the same functional region. An important contrast is that also imports bring technology spillovers as was recognized in Coe and Helpman, (1995), later followed by additional studies that avoid some of the econometric problems in the contribution by Coe and Helpman. The emerging understanding is that the more R&D intensive the imports are from other countries, the more can a region and a country accumulate of foreign R&D capital. Thus, import flows from countries with high R&D intensity seem to spur productivity growth in the importing country (and region) more than other imports. One may also distinguish import flows in general from imports of capital or equipment goods, and such studies indicate that the latter have a more distinct influence than overall imports. These different findings suggest that an individual firm benefits from knowledge embedded in import flows. Moreover, they suggest an advantage for functional urban regions in which firms collectively have rich and diversified imports from R&D and innovation-intensive origins. In these regions firms are positively stimulated in their renewal activities (Keller 2004; Andersson and Johansson 2010).

The basic idea of studies of technology spillovers is to find out how intra-firm (and intra-industry) R&D together with R&D of input suppliers combine to generate firm (and industry) *TFP growth*. As reported in Wolff (1997, 2012), the social rate of return to R&D is considerably larger than the direct *return to R&D*. These studies are frequently using industry level data and do not disentangle input-embodied innovation from knowledge flows in a more general sense.

21.5.3 Innovation Ideas and New Products

A product innovation leading to the marketing of a new good or service may have firms and/or households as major customer groups. Although the basic needs of consumers may be limited, there are myriads of changes occurring at the intermediate stages of production as well as in the individual choice processes of households. Regardless of whether we consider intermediate or final users, advancing sophistication and technological evolution consists mainly of substituting new means of consumer satisfaction for old ones. Under these circumstances the diffusion model has to be extended to take the form of a substitution model, recognizing

that the introduction of novelties generates the disappearance of established products. Ultimately new product attributes replace old attribute combinations (Batten and Johansson 1989).

Studies of knowledge flows and diffusion may focus on how such flows are more frequent and faster for certain types of product as well as process innovations, for certain types of firms, and/or for certain regional innovation milieus. Recent contributions emphasize that it is not enough to characterize firms and their capabilities and firm-specific networks. The innovation milieu and its knowledge flows play a fundamental role as innovativeness depends critically on a firm's possibility to combine internal and external knowledge resources.

A *product innovation* has to be marketed and commercialized. This part of the innovation effort brings us to product cycles that may be identified for a specific product group as well as for individual product varieties which belong to the same product group. As discussed earlier in this chapter, it is frequently claimed that product cycles with high frequency are initiated in metropolitan environments in which new product ideas are more prevalent and diverse than elsewhere. Large urban regions also host producer-service suppliers across a range of specialization, and these suppliers are important supporters in the commercialization process. As new products successfully penetrate domestic and foreign markets, they often reach a state of maturity in which knowledge intensity has a reduced role to play and many decomposed activities may be relocated to smaller urban regions or regions which have favorable factor prices for other reasons.

The diffusion phenomenon as described by Eq. (21.1) may be applied to depict a novel product's penetration of geographic markets. Penetration of this kind develops along an *S-shaped* curve like A in Fig. 21.3. For a given product or product variety, the exact curvature will vary with regard to trade links between a supply origin and relevant destinations.

The main message from Fig. 21.3 is that *market penetration* involves two interlinked processes, combining into a substitution process in which a novel product group B is substituted for an old one C. The oldest group in the figure is an aggregate of mechanical typewriter varieties which lose their joint market share as the new group of electrical typewriter varieties gradually gains an increasing market share. The third step is the simultaneous market share decline for electrical typewriters as these are replaced by word processing equipment. In addition, the reader already knows that the word processors rather quickly were replaced by personal computers (PCs).

21.6 Conclusions

Section 21.2 of this chapter has a discussion of the heterogeneity of firms. The conclusions will instead emphasize that functional urban regions are heterogeneous. However, we first have to stress that innovation is a firm activity and so is adoption of technical solutions. These activities rely on a firm's innovation strategy, its innovation efforts, and its renewal capabilities, where the strategy comprises

ambitions to develop capabilities, resources for innovation expenditures, and networks for knowledge flows.

The innovation milieu of the firm can be viewed as an innovation and R&D infrastructure, which facilitates the innovating firm's attempts to combine internal knowledge resources with knowledge sources in the firm's environment. The implicit suggestion of this chapter is that regions differ markedly in their supply of knowledge-intensive labor, knowledge exchange partners, knowledge-based producer services, and knowledge flows in general.

One may formulate a long list of characteristics of an urban region that make the region innovation supportive, providing the region's firms with favorable preconditions. With a shorter list, the regional innovation milieu can be advantageous in the following dimensions:

- The region can attract human resources (knowledge-intensive, creative, and talented individuals).
- The region can attract firms which benefit from access to knowledge sources and R&D activities in firms and universities in the region.
- The region can attract firms which are stimulated by an economic milieu where firms have extensive export networks and associated experiences.
- The region can attract firms that benefit from the regional presence of firms with well-developed import networks as well as import agencies and other firms specialized in selective import for local customers.

The enumerated (and related) characteristics create problems for empirical studies because of grave multicollinearity patterns. At the same time, they represent a particular form of "endogeneity": the composition of the firms which have the same region as a host constitutes the most essential attraction factor of the region. The result is a process of cumulative dynamics which maintain and develop favorable milieu characteristics in certain regions (in which the cumulative causation works in the desired direction), while the dynamics may cause milieu deterioration in other regions.

References

- Adams JD, Jaffe AB (1996) Bounding the effects of R&D: investigation using matched establishment-firm data. *Rand J Econ* 27(04):700–721
- Aghion P, Howitt P (1992) A model of growth through creative destruction. *Econometrica* 60(02):323–351
- Andersson M, Gråsjö U (2009) Spatial dependence and the representation of space in empirical models. *Ann Regional Sci* 43(1):159–180
- Andersson M, Johansson B (2010) Heterogeneous distributions of firms sustained by innovation dynamics – a model with empirical illustration and analysis. *J Ind Compet Trade*. doi:10.1007/s10842-010-0092-z
- Andersson M, Johansson B, Karlsson S, and Lööf H (2012) Introduction: Innovation and growth – from R&D strategies of innovating firms to economy-wide technological change. In: Andersson M, Johansson B, Karlsson C and Loof H, eds (2012), *Innovation and Growth – from R&D strategies of innovating firms to economy-wide technological change*, Oxford University Press, pp 1–20.

- Batten DF, Johansson B (1989) Dynamics of product substitution. In: Andersson ÅE, Batten DF, Johansson B, Nijkamp P (eds) *Advances in spatial theory and dynamics*. North-Holland, Amsterdam, pp 23–44
- Beckmann MJ (2000) Interurban knowledge networks. In: Batten D (ed) *Learning, innovation and urban evolution*. Kluwer, London, pp 127–135
- Capello R (2002) Entrepreneurship and spatial externalities: theory and measurement. *Ann Reg Sci* 36(3):387–402
- Coe DT, Helpman E (1995) International R&D spillovers. *Eur Econ Rev* 39(5):859–887
- Cohen WM, Levin RC (1989) Empirical studies of innovation and market structure. In: Schmalensee R, Willig R (eds) *Handbook of industrial organization*. North-Holland, Amsterdam
- Dosi G, Nelson RR (2010) Technical change and industrial dynamics as evolutionary processes. In: Hall BH, Rosenberg N (eds) *Economics of innovation*. Elsevier, Amsterdam, pp 381–410
- Fischer MM, Varga A (2003) Spatial knowledge spillovers and university research. *Ann Reg Sci* 37(2):303–322
- Geroski PA (1998) An applied econometrician's view of large company performance. *Rev Ind Organ* 13(3):271–294
- Griliches Z (1957) Hybrid corn: an exploration in the economics of technological change. *Econometrica* 25(4):501–522
- Hägerstrand T (1967) Innovation diffusion as a spatial process. The University of Chicago Press, Chicago
- Henderson V (1997) Externalities and industrial development. *J Urban Econ* 42(3):449–470
- Jacobs J (1969) *The economy of cities*. Random House, New York
- Johansson B, Karlsson C (1991) Från brukssamhällets exportnät till kunskapsamhällets innovationsnät (From the export networks of the natural resource economy to the innovation networks of the knowledge economy). *Länsstyrelsen i Värmlands län & Högskolan i Karlstad*
- Johansson B, Quigley J (2004) Agglomeration and networks in spatial economics. *Pap Reg Sci* 83(1):165–176
- Keller W (2004) International technology diffusion. *J Econ Lit* 42(3):752–782
- Klette TJ, Kortum S (2004) Innovating firms and aggregate innovation. *J Polit Econ* 112(5):986–1018
- Mansfield E (1968) *The economics of technological change*. Norton, New York
- Mansfield E (1987) Diffusion of technology. The new Palgrave a dictionary of economics. Palgrave Macmillan, Hampshire, pp 842–844
- Nyström K (2006) Entry and exit in Swedish industrial sectors. In: JIBS Dissertation Series No. 032, Jönköping International Business School
- OECD (2005) Oslo manual: guidelines for collecting and interpreting innovation data, 3rd edn. OECD, Paris
- Ohlin B (1933) *Interregional and international trade*. Harvard University Press, Cambridge, MA
- Polanyi M (1966) *The tacit dimension*. Doubleday, New York
- Romer PM (1986) Increasing returns and long-run growth. *J Polit Econ* 94(5):1002–1037
- Schumpeter JA (1934) *The theory of economic development*. Oxford University Press, New York
- Teece DJ (2010) Technological innovation and the theory of the firm. In: Hall B, Rosenberg N (eds) *Handbook of the economics of innovation*. North-Holland, Amsterdam, pp 679–730
- Vernon R (1966) International investment and international trade in the product cycle. *Q J Econ* 80(2):190–207
- Von Hippel E (1994) Sticky information and the locus of problem solving: implications for innovation. *Manag Sci* 40(4):429–439
- Wolff EN (1997) Spillovers, linkages and technical change. *Econ Syst Res* 9(1):9–23
- Wolff EN (2012) Spillover, linkages and productivity growth in the US economy, 1958 to 2007. In: Andersson M, Johansson B, Karlsson C and Loof H, eds (2012), *Innovation and Growth – from R&D strategies of innovating firms to economy-wide technological change*, Oxford University Press, pp 233–265.

Knowledge Flows, Knowledge Externalities, and Regional Economic Development

22

Charlie Karlsson and Urban Gråsjö

Contents

22.1	Introduction	414
22.2	The Nature of Knowledge	416
22.3	Knowledge Flows	418
22.4	Aspects of Spatial Knowledge Externalities	419
22.4.1	Economic Nature of Knowledge Externalities	420
22.4.2	Sources of Knowledge Externalities	421
22.4.3	The Recipients of Knowledge Externalities	422
22.4.4	Mechanisms of Knowledge Externalities	422
22.4.5	The Geographic Reach of Knowledge Externalities	424
22.4.6	Consequences of Knowledge Externalities	425
22.5	Knowledge Externalities and Regional Economic Development	426
22.5.1	New Economic Geography Models and Knowledge-Based Regional Growth	429
22.5.2	Interregional Knowledge Flows and Multiregional Growth	431
22.5.3	Microeconomic Aspects of Knowledge Spillovers, Knowledge Externalities, and Regional Economic Development	432
22.6	Knowledge Externalities and Regional Economic Development: Policy Conclusions	435
22.7	Conclusions	435
	References	436

C. Karlsson (✉)

Jönköping International Business School, Jönköping University, Jönköping, Sweden
e-mail: Charlie.Karlsson@jibs.hj.se

U. Gråsjö

Economics and Informatics, University West, Trollhättan, Sweden
e-mail: urban.grasjo@hv.se

Abstract

New knowledge generated by an economic agent in a region will tend over time to flow to other economic agents in the same region but also to economic agents in other regions. It is quite common in the literature to use the concept of knowledge spillovers for such knowledge flows, irrespective of whether they are intended or non-intended. The potential for intra-regional knowledge spill-over effects depends on the volume and character of the generation on new knowledge in each region as well as of the general characteristics of the individual regional economic milieu, that is, those location attributes, which are regionally trapped and which include how well integrated it is with other regions. The larger this potential, the higher the probability that firms dependent upon knowledge spillovers will locate there and the higher the probability that entrepreneurs will take advantage of this potential to launch innovations and to create new knowledge-based firms. To the extent that firms and entrepreneurs can enjoy these knowledge spillovers, they represent an externality or more specifically a knowledge externality in the regional economy.

Great importance is in the literature attributed to knowledge spillovers and knowledge externalities as drivers of regional economic development. Some authors, for example, claim that regional variations in localized knowledge spillovers are one of the main reasons behind regional variations in innovation performance. Against this background, the purpose of this chapter is, based upon a general characterization of knowledge flows, to analyze the character of knowledge externalities and, in particular, their sources, their economic nature, their recipients, their mechanisms and channels, their geographic reach, and their economic consequences generally and for regional economic development in particular.

22.1 Introduction

According to the new endogenous growth theory, cumulative processes characterized by either self-reinforcing decline or self-reinforcing growth drive regional economic development, which may last for long periods and which transform location patterns considerably. Knowledge accumulation is here a crucial factor for generating sustained regional economic growth. In the different functional regions, such dynamics are recognized as change processes involving location and migration of firms and households. Thus, a significant part of regional economic growth can be modeled as induced by changes in knowledge, technology, organization, and location, which are related to regional scale effects and durable, that is, slowly changing, regional characteristics. In particular, models of endogenous growth normally treat knowledge capital as an independent production factor, which, however, partly has the character of a public good. Hence, part of the new knowledge generated by an economic agent in a region will tend over time to flow to other economic agents in the same region but also to economic agents in other

regions. It is quite common in the literature to use the concept of knowledge spillovers for such knowledge flows, irrespective of whether they are intended or non-intended. A number of studies also provide evidence of the contribution of “knowledge spillovers” to economic growth. However, these studies have clear weaknesses. They, for example, do not clarify the mechanisms and channels through which knowledge spills over, that is, flows, and they assume that all economic agents equally benefit from the spillovers. However, to understand how knowledge flows affect regional economic growth and how regional policies can influence regional economic growth via measures that affect knowledge flows, we need to understand the role and function of different mechanisms and channels for knowledge flows, that is, we must apply a microeconomic perspective.

Knowledge can flow between economic agents (i) in the form of a knowledge transaction, where economic agents, for example, purchases a patent; (ii) as a by-product in connection with normal purchases of goods and services; and/or (iii) through other interaction between economic agents or their employees. When the knowledge flows are not fully compensated, we talk about knowledge spillovers. Griliches (1992) defined knowledge spillovers as “working on similar things and hence benefiting much from each other’s research,” but it is important to stress here that we do not restrict knowledge spillovers only to occur in connection with research, since knowledge also can be generated in other ways, such as “learning-by-doing.”

Intra-regional knowledge spillovers, that is, localized knowledge spillovers, may generate dynamically increasing returns in the regional economy and, thus, stimulate innovation and regional economic growth. Such increasing returns have long-term effects, since they create a tendency for any given ranking of the competitive positions of regions to persist over time. The potential for intra-regional knowledge spillover effects is a function of the volume and character of the generation on new knowledge in each region as well as of the general characteristics of the individual regional economic milieu, that is, those location attributes, which are regionally trapped and which include how well integrated it is with other regions. The larger this potential, the higher the probability that firms dependent upon knowledge spillovers will locate there and the higher the probability that entrepreneurs will take advantage of this potential to launch innovations and to create new knowledge-based firms. To the extent that firms and entrepreneurs can enjoy these knowledge spillovers, they represent an externality or more specifically a knowledge externality in the regional economy.

Great importance is in the literature attributed to knowledge spillovers and knowledge externalities as drivers of regional economic development. Some authors, for example, claim that regional variations in localized knowledge spillovers are one of the main reasons behind regional variations in innovation performance (Jaffe et al. 1993). Against this background, the purpose of this chapter is, based upon a general characterization of knowledge flows, to analyze the character of knowledge externalities and, in particular, their sources, their economic nature, their recipients, their mechanisms and channels, their geographic reach, and their economic consequences generally and for regional economic development in particular (cf. Johansson 2005).

This chapter is organized as follows: In [Sect. 22.2](#), we discuss the nature of knowledge and its effect on the economy. Knowledge flows are discussed in [Sect. 22.3](#). [Section 22.4](#) highlights aspects of knowledge externalities including their economic nature, the sources of knowledge externalities, the recipients of knowledge externalities, the mechanisms and channels of knowledge spillovers, the geographic reach of knowledge spillovers, and the consequences of knowledge spillovers. The relationships between knowledge externalities and regional economic development are discussed in [Sect. 22.5](#), which is followed up in [Sect. 22.6](#) with a discussion of policy conclusions. [Section 22.7](#) concludes.

22.2 The Nature of Knowledge

In order to discuss knowledge flows and the knowledge spillovers and externalities they may generate, we must first discuss the nature of knowledge and its effects on the economy. A natural starting point is to examine how to distinguish between information and knowledge. Information can be defined as messages or data, which can easily be codified and therefore transmitted, received, transferred, and stored at low costs. Of course, these messages can communicate statements about knowledge. Knowledge, on the other hand, consists of organized or structured information that is difficult to codify and interpret and thus to transform into useful messages, generally due to its intrinsic complexity and indivisibility. Thus, much knowledge is tacit because it is the cumulative output of long periods of learning, specific to a particular setting, and cannot easily be written down and become codified.

Focusing on knowledge that can be related to the activities of economic agents, we can distinguish between the following categories:

- Know-how, which is always embodied in persons or embedded in economic agents, such as firms and other organizations. It signifies expertise, skills, and practical attainments. Know-how can be present without codified instructions, generically based on experience and training and often so difficult (or uneconomical) to codify that it remains tacit.
- Know-why, which has the character of systematic and broadly accepted (scientific) explanations, which can be stored in codified form but which may require specific training and skills to be decoded and understood. It refers to a capacity to understand and explain. Know-why refers to science in the sense that it does not exist – by definition – if it has not been created and codified.
- Knowledge in the form of human capital, which represents a combination of know-how and know-why embodied in persons.
- Knowledge embodied in products (hardware and software) created by persons or economic agents by applying human capital, know-why, and know-how in some production process.

For knowledge spillovers and related externalities, it is essential to consider the degree to which knowledge is “rivalrous” and “excludable” (cf. Cornes and Sandler [1986](#)). A rival good has the property that its use by one economic agent precludes

the use by another economic actor, whereas a non-rival good lacks this property. Excludability relates to both technology and legal systems. A good is excludable if the owner can prevent others from using it. Pure public goods are both non-rival and non-excludable. This creates a fundamental conflict in society, since an economic agent will only be motivated to carry out R&D if competitors can be excluded, whereas society will benefit if the knowledge (innovation) is allowed to diffuse to be used by many economic agents (cf. Arrow 1962).

What types of knowledge are then relevant from the perspective of an economic agent? The following primary types of knowledge can be considered:

- Knowledge about activity routines
- Knowledge about output varieties
- Knowledge about markets and customers' willingness to pay for different output varieties
- Knowledge about routines to develop activities and output development activities

Activity routines include techniques, methods, and approaches that to a varying extent are applied in production, administration, distribution, logistics, transaction, interaction activities, and for innovative economic agents also activity and output development activities including general search for new knowledge (cf. Nelson and Winter 1982). The routines are a manifestation of the know-how of an economic agent. From a related perspective, it is possible to identify the following three knowledge concepts: (i) scientific knowledge (principles), (ii) technological knowledge (blueprints), and (iii) entrepreneurial (business) knowledge. In this context, it seems important to remark that both scientists and engineers perform R&D activities, while making use of know-how about effective and feasible ways to conduct research. Attempts to codify such know-how are often quite primitive and superficial.

However, whatever knowledge concept we use, it is important to observe that knowledge is distributed across a large number of individuals and other economic agents in every economy and that this distributed knowledge must be combined in new ways to generate new knowledge and innovations. Thus, to generate new knowledge, economic agents are dependent upon both "inside" and "outside" knowledge, where the former refers to an economic agent's own investments in new knowledge, such as R&D, whereas the latter refers to knowledge production activities performed by other economic agents. Own investments in knowledge production can, from this perspective, be comprehended as a means to absorb and appropriate "outside" knowledge (Cohen and Levinthal 1989). This absorptive capacity is also a function of the employment of skilled and highly educated individuals by economic agents, since the presence of such employees seems to be a key channel by which knowledge is transmitted across economic agents.

Understanding how combinations of distributed knowledge generate new knowledge demands interactive models of knowledge creation and innovation which can be achieved within national, regional, and metropolitan systems of innovation. This implies that new knowledge is not created in some anonymous knowledge production process. Instead, new knowledge is the result of interaction between often identifiable individuals, who previously have accumulated

a substantial stock of knowledge in their specific fields of expertise but who also more or less constantly are keeping themselves updated through various knowledge channels to be aware of new knowledge created elsewhere. New knowledge is created when these individuals share their knowledge within a larger group of people, for example, at a university department or in a research institute or in the research department of an economic agent.

Once it is recognized that new knowledge and new ideas are often offsprings, variations, or combinations of existing knowledge, the significance of “outside” knowledge becomes clear. As knowledge as an input in knowledge production is non-rival and tends to diffuse and spill over despite various appropriability mechanisms, it is frequently maintained that there are increasing returns in knowledge production. For such returns to emerge, diffusion and spillovers of the accumulated and, in particular, the newly generated knowledge are central, as it must be accessible for other economic agents as an input in their knowledge production. However, we must acknowledge that knowledge is often extremely complicated and contains complex elements. This implies that it often only is accessible via interactions within either the economic agents’ innovation networks or general innovation systems that according to much of the literature in the field tend to be bounded by geographical proximity (Jaffe et al. 1993; Karlsson and Manduchi 2001).

22.3 Knowledge Flows

Knowledge flows can be described as a special sort of communication related to the diffusion of messages, products, individuals, or economic agents that embody new ideas, knowledge, concepts, blueprints, and so on (Rogers 1983). Such flows occur whenever an idea generated by a certain economic agent is learned by another economic agent and indicate a process where economic agents learn from another economic agent’s ideas and combine these with internally generated ideas and internally existing ideas, thereby developing and extending the internally existing stock of ideas (cf. Griliches 1992). However, this learning can occur through many different mechanisms, such as markets, publications, social networks, professional networks, education and training, and labor mobility, which indicates that the diffusion of knowledge is a complex matter to disentangle and to understand not least since it is also dependent upon formal and informal institutions and the level of social capital (Helpman 2008, Ed.). Concerns have been raised that, on the one hand, the role of knowledge diffusion is underestimated and that, on the other hand, the role of knowledge spillovers, that is, unpriced knowledge externalities, compared to normal market transactions of knowledge is overestimated in economic theory (Breschi and Lissoni 2001).

As the creation of knowledge is spatially concentrated, it is obvious that knowledge flows that diffuse knowledge spatially play a decisive role in regional economic development. Due to its character, a substantial part of the diffusion of more complex knowledge takes place through face-to-face interaction. Frequent face-to-face

interaction brings distinct information including persistent updates, planned and unplanned learning and the development of similar interpretation schemes, shared understanding of new knowledge and technologies, local institutions, and similar cultural traditions and habits. Developments within “cognitive science” have since several decades credited processes of face-to-face interaction a fundamentally important role for knowledge accumulation and knowledge generation among economic agents. Spatial proximity facilitates face-to-face interaction and, thus, tends to accelerate the transfer of knowledge between individuals and other economic agents. However, spatial proximity is not a sufficient condition for knowledge transfers to take place. Learning often requires trust and cognitive and social proximity, which of course may be facilitated by spatial proximity.

There is actually a long tradition in economics studying these knowledge flows in terms of knowledge diffusion and knowledge spillovers. However, it is in particular since pioneering work of Griliches (1992) that knowledge flows have been extensively analyzed in the micro-productivity literature but also in a growing literature on knowledge spillovers in a spatial context.

The study of knowledge flows has used a number of distinct approaches and techniques. What can be characterized as a technological approach has assumed that knowledge only flows between economic agents (firms) in the same technology group. Other studies use more innovative measures of knowledge flows and define, for example, technological distance as a bilateral measure, which permits different flow intensities between different pairs of economic agents (Jaffe 1986).

Since the early 1990s, researchers have increasingly used patent citations to follow the path of learning. Patent citations in fact report the potential learning flows between the citing and the cited economic agent. The importance of distance for knowledge flows has been tested by means of patent data (Jaffe et al. 1993). Patent data has also been used for analyzing and comparing knowledge flows originating in universities, federal labs, and firms (Jaffe 1989).

22.4 Aspects of Spatial Knowledge Externalities

Knowledge externalities can occur in situations where the protection of proprietary knowledge is incomplete. To be able to disentangle the role of knowledge externalities in regional economic development, we need to distinguish between six aspects of knowledge externalities: (i) their economic nature, (ii) their sources, (iii) their recipients, (iv) their mechanisms, (v) their geographical reach, and (vi) their effects (cf. Johansson 2005). Knowledge externalities involve firms but also other types of organization, and to have a more general discussion, we use the more general term economic agents. Economic agents are characterized by two main types of activities:

- Ordinary production activities, that is, at each point in time an economic agent uses current and fixed inputs to produce output by means of given techniques (routines)

- Development activities, that is, the use by economic agents of part of the inputs to develop new types of outputs and/or new routines (including the development routines)

It is essential to distinguish between these two types of activities, since knowledge externalities have quite different effects in the two cases. Both ordinary production activities and development activities involve interaction with other economic agents – interactions that give rise to interaction costs, which increase with geographical distance between the actual economic agents and which are nonlinear with regard to geographical distance. Hence, proximity brings an advantage to economic agents.

22.4.1 Economic Nature of Knowledge Externalities

As regards the economic nature of knowledge externalities, we have:

- Pecuniary knowledge externalities that operate via prices, that is, via market links (intra-market knowledge externalities) or via interorganization links (quasi-market knowledge externalities), which we may term connected knowledge spillovers
- Nonpecuniary (technological) knowledge externalities, which operate outside the market, that is, extra-market knowledge externalities or if we like pure knowledge spillovers

In perfect markets, there are no contacts between economic agents. In principle, they do not know each other. Some markets are rather close to perfect markets. However, many markets are characterized by different forms of links between economic agents. A suitable starting point for understanding such markets is the microlevel of individual decision-making units. The decision-making unit can be an economic agent in the form of a firm or other organization or household or individual decision-makers within firms and other organizations. A basic presumption here is that firms and other organizations have internal networks for communication and coordination of production and development activities. Certain internal networks consist of links that are arranged for the flow of resources, while other internal networks function as channels for diffusion and exchange of information and knowledge. All internal networks are connected and governed in such a way that economic agents are coherent.

Attached to the internal networks of economic agents are links that extend beyond the boundaries of the economic agent. Such links connect various economic agents with each other and constitute what we can call interorganizational networks, which are used for flows of goods, services, and/or knowledge. Interaction in this kind of knowledge networks can lead to the development of hybrid forms of knowledge that are freely available only to the network members and thus neither public nor private in character. This hybrid knowledge becomes a kind of club good for the network members (cf. Buchanan 1965). Interaction between economic agents is based upon a formal or informal contract, which normally is long term if one or several of the economic agents involved must make investments that are transaction or link specific.

22.4.2 Sources of Knowledge Externalities

New knowledge is generated by economic agents through deliberate search for new knowledge in the form of R&D activities and through learning-by-doing when carrying through different activities. This search is directed toward different kinds of sources. The sources can be classified into two groups: (i) sources containing embodied knowledge, including individuals, economic agents, and products, and (ii) sources containing disembodied knowledge, including books, articles, research and consultancy reports, patents, and web pages. Since the knowledge-generating activities are localized, regional knowledge externalities have two main spatial sources: (i) intra-regional knowledge sources, that is, knowledge sources characterized by geographical proximity, which lower transaction costs and facilitates pure knowledge spillovers, and (ii) interregional knowledge sources, that is, knowledge sources in other regions, available via different interregional links. Proximity implies that transactions and planned interactions between economic agents become less costly and that the probability for non-planned interactions increases. More advanced types of interactions between economic agents are dependent upon trust, and proximity makes it easier to develop trust (Breschi and Lissoni 2001).

However, for proximity to be important, it must be proximity to something, in this case other economic agents with the relevant knowledge. One type of proximity is proximity to economic agents engaged in the same trade. Such clustering or colocation gives rise to a special form of agglomeration economies, namely, localization economies, where specialization is an important feature. Clustering or colocation of economic agents from different trades gives rise to another type of agglomeration economies, namely, urbanization economies, where diversity and size are essential features. Agglomeration implies that economic agents can benefit from mutual proximity, but whether a more specialized or a more diversified regional economic milieu is most favorable for existence and size of localized knowledge spillovers is still an open question. It is in this connection important to observe that knowledge generation activities are more highly agglomerated than most other types of economic activities.

Anyhow, even if proximity is no guarantee for knowledge spillovers between economic agents, it has two important potential effects. It affects both how economic agents can interact in the marketplace and how they can interact outside the marketplace and benefit from nonpecuniary knowledge spillovers. Thus, proximity influences intra-market and quasi-market as well as extra-market externalities. However, quasi-market externalities are different, since they represent a link between economic agents. Establishing a link to another economic agent is a means to reduce geographical interaction costs and thus to make proximity less critical, that is, links can function as a substitute for proximity. However, since the establishment of an economic link involves search and negotiation costs, we must consider that link formation within a region can be less costly than formation of a link to an economic agent in another region, which implies that proximity play a role for extra-market externalities also.

22.4.3 The Recipients of Knowledge Externalities

Knowledge spillovers across economic agents occur when the knowledge generated by one economic agent is “borrowed” by other economic agents. Here, we must distinguish between spillovers between economic agents in the same trade and spillovers between economic agents in different trades (Audretsch and Feldman 1999). One critical question here concerns whether the specific mix of economic activities undertaken within different regions matter for the extent and direction of knowledge spillovers, that is, do knowledge spillovers occur mainly within or between trades? This question, which concerns the recipients of knowledge spillovers, is very relevant, and a debate among researchers during the two last decades has focused precisely on how the knowledge externalities generated by knowledge spillovers are affected by the regional mix of economic activities. Despite a consensus that knowledge spillovers within a given region stimulate dynamic knowledge externalities, there is no agreement concerning which the recipients of these knowledge spillovers are.

Glaeser et al. (1992), which analyze the factors that influence innovative activities in urban regions, illustrate the controversy. The authors identify two relevant models in the economics literature. The first model is the so-called Marshall-Arrow-Romer (MAR) model, which formalizes the insight that the concentration of a particular trade within a specific urban region (Lösch 1954) promotes intra-regional knowledge spillovers across economic agents in that particular trade and therefore stimulates innovation in that particular industry. The basic assumption here is that knowledge spillovers, and thus knowledge externalities, mainly takes place across economic agents in the same trade.

The alternative view regards intertrade knowledge spillovers as the most important channel to diffuse new economically relevant knowledge. Not least, Jacobs (1969) argues that the agglomeration of economic agents from different trades in urban regions fosters innovations due to the diversity of knowledge sources located in such regions. The recipients of knowledge spillovers from economic agents in one particular trade are here economic agents in other trades. The assumption here is that the variety of industries within an urban region can be a powerful engine of growth for that region and that the exchange of complementary knowledge across diverse economic agents leads to increasing returns to new knowledge. The empirical studies in the field give no clear answer to the question whether MAR or Jacobs’s externalities are most important.

What must be observed is that the gains from knowledge spillovers do not apply uniformly across the economic agents in a region due to the heterogeneity among economic agents. They differ in terms of their history, age, size, knowledge and other resources, location, networks, ownership structure, routines, strategies, and behavior even if they belong to the same industry.

22.4.4 Mechanisms of Knowledge Externalities

Breschi and Lissoni (2001) argue that it is important to improve the understanding of the knowledge transmission mechanisms in addition to study knowledge

spillovers by a rather limited set of variables. The mechanisms conveying knowledge externalities include (i) formal and informal interaction between economic agents, where the formal interaction is based on an explicit contract, while the informal interaction is based upon an implicit contract; (ii) active knowledge search of economic agents; and (iii) mobility of economic agents.

The formal and informal interaction between economic agents can take many forms ranging from transactions of goods and services including R&D services to cooperation in the form of joint ventures and strategic alliances including R&D cooperation. It includes the interaction of employees of different economic agents privately and in social, civic, and professional organizations. Since much knowledge is embodied in people, it is natural to assume that knowledge spillovers are partly a function of the interaction between people with the relevant education, skills, and experiences. Relations to suppliers and/or customers are also potential channels for knowledge spillovers. Furthermore, trade with goods and services embodying knowledge is a further channel for knowledge spillovers and externalities (Grossman and Helpman 1991b). Lastly, we have public policy programs supporting the cooperation between economic agents, which can be potentially important channel for knowledge spillovers.

Active knowledge search involves what can be described as “business intelligence” and involves activities ranging from analyses of patent applications and academic publications to “reverse engineering” of products. Mobility of economic agents involves the mobility of labor as well as the mobility of firms and other organizations. The mobility of people between economic agents and between regions is a potentially important channel for knowledge spillovers. We may observe that there exist several more mechanisms, which support and facilitate the transfer and diffusion of tacit as well as codified knowledge and technology: (i) education; (ii) seminars, conferences, and trade fairs; (iii) interactive communication channels (E-mail, the Internet, video conferences, etc.); (iv) people specially designated to obtain and disseminate knowledge (e.g., gatekeepers); (v) knowledge management within and between economic agents; and (vi) imitation.

It is important to notice that even if each of these channels or mechanisms can be seen as partly independent of each other, they are often linked to each other in different ways. It is in this connection important to observe that international cooperation in both the private and the public sectors play an important role for knowledge diffusion. An increasing number of partnerships among firms, universities, and public research centers as well as between individual researchers and inventors is a clear indication of the growing importance of collaboration. Collaboration permits the partners to share and acquire the expertise of each other, thus enriching their overall know-how. It often functions as a positive sum game, where the advantages outweigh the disadvantages even if the advantages are not always equally shared among the partners.

We are now in a position where we can disentangle the mechanisms behind intra-market, quasi-market, and extra-market knowledge externalities. Starting with the intra-market knowledge externalities, it is obvious that location and urbanization economies will make it possible for economic agents to buy inputs embodying

knowledge with lower transaction costs and potentially also at a lower price due to proximity. We can here think of hardware as well as software embodying knowledge, but we can also think of knowledge-intensive business services as relevant here. In particular, we can here think of purchases that are made ad hoc or so seldom that it is not rational to invest in a link. Urbanization economies bring diversity as an extra dimension, which gives economic agents an opportunity to test different suppliers with a somewhat different knowledge base for the same type of basic delivery and generate what is known as Jacobs's externalities (Jacobs 1969). Localization and urbanization economies also make it possible for economic agents to sell knowledge-intensive outputs with lower transaction costs and potentially at a lower price due to proximity.

Quasi-market or transaction-link knowledge externalities represent idiosyncratic relations between economic agents that provide the participants with advantages that occur in a quasi-market setting besides the ordinary market, that is, these knowledge externalities are a kind of club good. A link between economic agents for the delivery of knowledge or knowledge-intensive goods and/or services reduces transaction costs.

Extra-market knowledge externalities concern information and knowledge spillovers that occur as by-products in the course of all types of interactions between economic agents but can also occur without any direct interaction between economic agents. The prime focus here is on research, development, and innovation activities that are assumed to be stimulated by such spillovers (Karlsson and Manduchi 2001).

However, we must also acknowledge that knowledge can spillover between economic agents without any direct interaction. Obvious examples are when economic agents analyze patent applications and academic publications to get inputs, in particular, to their own research, development, and innovation activities.

22.4.5 The Geographic Reach of Knowledge Externalities

A critical issue in analyzing the role of knowledge spillovers and thus knowledge externalities for regional economic development is the geographic or spatial reach of knowledge spillovers. We have many reasons to believe that knowledge is subject to spatial decay. Due to "the tyranny of distance," most human interaction takes place within the functional region and in particular the locality where people live and often work. The claim that geographical proximity matters for knowledge spillovers between economic agents is largely supported by the empirical literature (Karlsson and Manduchi 2001). Already Glaeser et al. (1992, 1127) maintain that spatial proximity facilitates knowledge spillovers, because "intellectual breakthroughs must cross hallways and streets more easily than oceans and continents." This is followed up by Audretsch and Feldman (1999, 410) who argue, "knowledge spillovers not only generate externalities, but the evidence suggest that such knowledge spillovers tend to be geographically bounded."

The statement by Audretsch and Feldman that knowledge spillovers "tend to be" geographically bounded indicates that knowledge spillovers also may occur

between regions. Actually, the authors themselves only 5 years later argued, “there is no reason that knowledge stop spilling over just because of borders, such as a city limit, state limit or national boundary” (Audretsch and Feldman 2004, 6). In this connection, it may be relevant to go back to Palander (1935), who observed that one of the most remarkable features of modern urban structures is the frequency and extension of the interactions between activities carried out in different cities. These interactions presuppose of course the possibility of communicating between cities. Possibilities that have multiplied many times since the 1930s due to, on the one hand, a telecommunications revolution that has lowered the marginal cost of information exchange between different locations to levels very close to zero and, on the other hand, the evolution of highway and air travel networks that significantly has reduced the travel costs and the travel times. Thus, the interregional interaction costs have been reduced substantially in recent decades creating the necessary foundations for a global knowledge-intensive network economy. Against this background, we may ask to what extent it actually is true that interregional knowledge spillovers are limited in scope and spatial reach.

If we first turn our intention to scientific knowledge, we can indeed claim that the interregional knowledge spillovers are both substantial and rapid between individuals with the relevant absorptive capacity. The reason for this is that the international scientific community is organized in big knowledge networks, relying, for example, on international scientific conferences and journals, and that rapid publication of new scientific results are important for the prestige of the individual scientists. Is technological knowledge then different? To a certain extent, it is, since the economic agent that developed new technological knowledge normally does not want to share it without compensation with competitors. On the other hand, the economic agent who has been able to develop an innovation based upon own-developed new technology is eager that it should be diffused rapidly among customers wherever they are located before any competitor imitates the innovation. Entrepreneurship knowledge in the form of business ideas probably diffuses between regions without major problems, since it normally is not proprietary, even if trademarks and logotypes can be protected. Thus, it might be the case that the claims that geographical knowledge spillovers tend to be geographically bounded underestimate the geographical reach of knowledge spillovers. Actually, there is abundant evidence that the information and knowledge networks that enhance the efficiency and innovativeness of economic agents can be and often are widely diffused geographically. However, we need to stress that a critical factor that we do not discuss here is the speed with which that knowledge diffuses between regions.

22.4.6 Consequences of Knowledge Externalities

The consequences of knowledge externalities appear in the following forms:

- Efficiency externalities, which generate static differences between regions with regard to productivity and unit costs of economic agents

- Innovation externalities, which are dynamic phenomena and appear as new knowledge inducing changing economic efficiency (new routines) but also in the form of new products, increased product characteristic diversity, and similar novelties

We may assume that intra-market knowledge externalities and thus proximity are of special importance for efficiency externalities (cf., Johansson 2005). However, we must observe that links, that is, quasi-market externalities can function as an alternative to proximity.

However, the focus here is mainly on innovation externalities. In recent decades, we have witnessed a substantial progress in the understanding of innovation processes both with regard to process and product innovations. In particular, this is true for the macro level, with the development of the endogenous growth models (Romer 1986; Lucas 1988). In these models, development of technology is an endogenous part of each model, and this endogeneity is coupled to assumptions about knowledge externalities.

Unfortunately, we must admit that the micro foundations for understanding innovation processes are less well developed. The micro-oriented research on innovation processes has had a much stronger empirical focus and has partly been centered on concepts such as regional innovation systems and innovative milieu.

22.5 Knowledge Externalities and Regional Economic Development

In this section, we turn our focus to the effects of knowledge externalities on regional economic development. How does knowledge flows generally influence regional economic development? Orthodox economic theory gives two answers:

- a. Knowledge affects regional economic development via the production functions of regional economic agents by improving their use of their inputs, that is, their productivity (Chambers 1988).
- b. Knowledge affects regional economic development via the value ladder of product varieties produced by regional economic agents (Grossman and Helpman 1991a).

However, orthodox economic theory says nothing about the knowledge generation and diffusion mechanisms in general and knowledge spillovers in particular. A natural starting point for an increased understanding of these issues is a simple endogenous regional growth model for an isolated region, since concepts related to knowledge generation, knowledge accumulation, knowledge appropriation, knowledge flows, and knowledge spillovers are prominent features of such models as well as in the literature on innovation systems (Lundvall 1995, Ed.). These models also provide a systematic approach to understanding the adaptive capacity of a regional economy. Endogenous growth models describe the growth process of the isolated region and suggest that continuous increases in knowledge (“technology”) due to investments in knowledge generation increase aggregated economic growth in the region (cf. Romer 1986, 1990). The basic idea behind such models is that part of

a region's resources is used to produce an output that can be used for consumption and investment, while the remaining resources are employed in producing new knowledge ("technology"). Since knowledge is treated as non-rivalrous and only partially excludable, these models exhibit increasing returns. However, imperfect competition is needed for R&D investments to be worthwhile for economic agents. The innovations generated through the R&D investments later become the intermediate inputs for other firms, and hence, the rate of innovation determines the overall rate of growth.

New (technological) knowledge is in this kind of models used in two ways in the regional economy:

- a. The economic agent in the region that developed it uses new knowledge in the production of a specific unique product. Other economic agents in the region are excluded from using the same knowledge by means of patenting.
- b. New knowledge increases the total stock of knowledge in the region but may spill over to other economic agents to be used in knowledge production by means of examination of patent documentation (Romer 1990). Thus, the knowledge production productivity in the region increases. It may very well be that the new knowledge benefits other economic agents as much or even more than the economic agent that created it.

One limiting factor of this original Romer approach is the assumption of general accessibility of the stock of knowledge in general and new knowledge in particular for all economic agents in the region. There are strong reasons to believe that in particular, new knowledge is not evenly accessible for economic agents in a region, since, for example, not all new knowledge is patented but instead kept as business secret and not all economic agents have the absorption capacity necessary to use the new knowledge. Another limiting factor is that knowledge is treated as a homogenous concept and that no distinction is made between different types of knowledge.

Due to their aggregated character, these models have limitations when it comes to understanding the relation between knowledge production and economic growth at different levels of identification: the level of economic agents, the level of industries, and the regional level. Knowledge is assumed to spillover between economic agents in the region and to generate knowledge externalities, but the precise mechanisms are not explicated.

Another problem with the single-region Romer model is the separation from all other regions. Economic agents trade with economic agents in other regions, and economic agents make direct investments in other regions. Furthermore, people travel and move between regions, so there are numerous mechanisms through which knowledge will diffuse from one region to another and have impact upon productivity and innovation and thus regional economic development in other regions.

An alternative type of endogenous growth model is the intentional human capital model, which stresses the critical importance of education, learning-by-doing, and knowledge spillovers. Technological progress is here the result of intentional research and education investments, and here, human capital is an element in the

aggregate production function. Here, investments in human capital generate knowledge spillovers, which increase the productivity of both physical capital and the general labor force.

Moving from a single-region to a multiregion framework leads to a number of complications, when we want to disentangle the extent to which knowledge externalities influence regional economic development. We must acknowledge:

- That regions are different not least in terms of agglomeration, industrial structure, and accessibility to and interaction with other regions
- That regional economic development is path dependent, that is, history matters
- That the capacity to adapt to economic, technological, and institutional changes varies between regions
- That self-organization prevails due to the actions of households and economic agents
- That knowledge and knowledge generation is not evenly dispersed but instead unevenly distributed and in particular concentrated in (large) cities
- That the distribution pattern differs for different types of knowledge and institutions

Thus, the development potential of different regions is dependent upon their historical development paths including the partly self-organized (dis-)accumulation of infrastructure, economic agents, knowledge, and institutions and the pertinent potential for intra- and interregional knowledge spillovers, which defines the potential development trajectory for each region. Certainly, those regions that offer larger knowledge and other externalities have *ceteris paribus* a larger development potential than other regions. The driver behind regional economic development is the initiation of new technologies, new product cycles, new industries, new infrastructures, and new institutions that over time complements or replaces old ones through processes with different speeds. Hence, the economic development in different regions hinges on their capacity to absorb, generate, and adapt knowledge and to generate and nurture such novelties. This capacity is, given our discussion above, dependent upon (i) the education level of the regional labor force and the incentives for individuals to participate in education, (ii) the incentives for economic agents to undertake R&D and to innovate, and (iii) the degree to which knowledge diffuses within the region and from other regions.

Endogenous growth models tell us that the regions that are most well equipped with resources are most likely to accumulate more resources and increase productivity more over time according to the principle of circular and cumulative causation (Myrdal 1957). However, this form of positive feedbacks is in general constrained, on the one hand, by the development of the demand of the region and its external markets and, on the other hand, by the existing capacities in the form of built environment, accessibility-based transport systems, production capacities, and labor supply. Anyhow, differences between regions may “accrue from the growth of industry itself – the development of skill and know-how; the opportunities for easy communication of ideas and experience; the opportunity of ever-increasing differentiation and of specialization in human activities” (Kaldor 1970, 340).

It is certainly beyond the purpose of this chapter to present a formal multiregional model that in a consistent manner integrates the effects of knowledge flows and pertinent knowledge spillovers on regional economic development. The discussion here is limited to a discussion of some of the factors that must be considered when modeling knowledge-based multiregional growth. We start with discussing new economic geography models and knowledge-based regional growth and continue then with discussing first the role of interregional knowledge on multiregional growth and then how knowledge flows influence the behavior of individual economic agents. New economic geography models combine economies of scale with reduced transport costs to explain why similar economic activities will tend to concentrate in the same locations. These locations will become densely populated and will also have higher levels of income (Krugman 1991).

22.5.1 New Economic Geography Models and Knowledge-Based Regional Growth

It is well known that the original new economic geography models are not suitable to model the role of knowledge spillovers and knowledge externalities in regional economic growth because Krugman originally refused to model technological externalities (Krugman 1991). However, the literature on innovation systems convincingly shows that:

- Knowledge generation is highly geographically concentrated and thus that regions differ substantially in their volume of knowledge production.
- Regions differ substantially in terms of absorption capacity and accessibility to external knowledge.
- Its own unique knowledge mix characterizes each region.

It also stresses that knowledge flows and pertinent knowledge spillovers are at the core of knowledge-based regional development. This implies that the location of knowledge generation and the structure of the spatial networks for knowledge flows and knowledge spillovers are fundamental factors in modeling knowledge-based regional economic development. In addition, it must be acknowledged that the potential of economic agents to absorb new scientific, technological, and entrepreneurial knowledge is facilitated by geographic proximity (Jaffe et al. 1993). The exchange of knowledge and ideas may generate such technological externalities that it influences the location decisions of economic agents and induces them to cluster at specific locations (Marshall 1920), generating differences in income and productivity across geographic space (Henderson 1974). It was first with the merger of new economic geography models and endogenous growth models that knowledge spillovers were taken into account in the new economic geography literature. The critical message here is that when human capital and other capital resources, including R&D competence, are located together in the same region, self-reinforcing processes can be obtained, that is, such concentrations have the power to further stimulate economic growth in these regions.

Large (and dense) regions offer special advantages in terms of knowledge flows and potential for knowledge spillovers, since they normally host research universities, private R&D facilities, and many industrial clusters, that is, industrial diversity. Small (and sparsely populated) regions on the other hand are normally characterized by a lack of research universities and private R&D facilities as well as by having only one or a few industrial clusters or being dependent upon large-scale industries. This suggests the possibility of formulating a model along the lines of new economic geography principles, which takes one of its starting points the favorable conditions for knowledge generation and large potential for knowledge flows and knowledge spillovers in large regions. When such a region for any reason has achieved an initial advantage in knowledge generation, it will attract knowledge-generating and knowledge-utilizing economic agents, since it offers opportunities to get advantages in terms of increasing returns in knowledge generation and knowledge utilization.

Although it is indeed true that it is economic agents and not regions that compete, there are important regional production and knowledge advantages being pursued by the many economic actors within large regions. This leads to an important premise about scientific, technological, and entrepreneurial knowledge and innovations in large regions. New knowledge and innovations emerge uniquely out of large regions not simply because one or the other was endowed with a certain initial stock of knowledge and factors of production but because many assets necessary for knowledge generation and innovations are created in the course of the ongoing knowledge generation, innovation, and production activities among economic agents. These assets include new scientific, technological, and entrepreneurial knowledge, which are prerequisites for novelty-by-combination processes, which generate new knowledge and innovations but which facilitate the exploitation of knowledge resources and the development of regional formal and informal institutional innovations to support and sustain intra-regional knowledge flows. Furthermore, many of these scientific, technological, entrepreneurial, and institutional assets are not easily transferrable between regions and thus may preserve regional advantages for a long time.

The size of the regional market potential influences the probability that new knowledge in the form of inventions is turned into innovations in the region. The underlying reason is that a large market potential generally increases the demand for knowledge-intensive products. The probability that inventions are turned into innovations increases with the size of the regional market potential, and this gives knowledge-creating and knowledge-utilizing economic agents an extra advantage of locating in large regions. In addition, when more of these economic agents locate in large regions, this makes these regions more attractive for knowledge workers, which fuels the cumulative process and makes the region still more attractive for knowledge-intensive economic agents. More knowledge handlers and more knowledge-intensive economic agents in a region increase its market potential and make it still more attractive. Thus, we may conclude that from an endogenous growth perspective, regions that have a first-mover advantage in terms of knowledge, technology, and innovation are likely to attract educated labor (and capital) from

other regions and thereby generating a cumulative self-reinforcing process of R&D and innovation leadership. This implies that the scarce and critical resources in knowledge-based regions are factors such as highly educated labor, education system, R&D system, knowledge flows and spillovers, and innovation capability.

When large innovative regions grow, the demand for land, premises, and labor increases, which tends to drive up their prices. However, these price increases are a necessary mechanism to secure that resources gradually are pressed out of maturing industries and made available for new growth industries. This implies that economic growth in large innovative regions is strongly related to the functioning of the relevant markets. Of course, a condition for long-term growth is that the advantages in terms of knowledge and other spillovers increase with at least the same speed as the agglomeration costs.

22.5.2 Interregional Knowledge Flows and Multiregional Growth

The discussion in this section builds upon an endogenous growth model introduced by Andersson and Mantsinen (1980). The model is centered on a production function for each region r , $Q_r = Q(K_r, A_r)$, where K_r represents the total production capital in region r and A_r the total accessible knowledge in region r . The knowledge resources in any region r are G_r , but region r can also benefit from knowledge resources in all other region s , G_s . Thus, A_r represents a compound knowledge consisting of knowledge within region r and accessible knowledge in other regions.

With the help of the above model components, it is possible in a stylized way to illustrate how regional and interregional knowledge resources influence regional economic growth. G represents here all types of knowledge as well as the capacity to develop new knowledge. Now, assume that knowledge can be treated as a spatial public good, such that G_s in any region s influences the accessible knowledge in region r via a distance-decay factor $f_{sr} = \exp\{-\lambda t_{sr}\}$, where λ is the travel time sensitivity and t_{sr} is the travel time distance between region r and region s . This implies that

$$A_r = \sum_s f_{sr} G_s$$

summarizing the total accessible knowledge in region r , where $f_{rr} > f_{sr} > 0$, for $s \neq r$. In this model, the accessible knowledge will change when any of the G -variables, travel sensitivities, and/or travel time distances change. Changes in the G -variables are the outcomes of knowledge generation processes, which can be illustrated by the following differential equation:

$$\dot{G}_r = H[g_r(A_r), \tau_r(Q_r), \theta_r(Q_r)]$$

where g_r is the R&D productivity in region r , which is a function of accessible knowledge in region r , A_r , τ_r is the share of output Q_r devoted to R&D in region

r and θ_r is the “learning-by-doing productivity” in region r , which is a function of total output in region r . The important message here is that regions do not only rely of their own internal knowledge and capacity to generate knowledge but also on their capacity to integrate knowledge generated in other regions. If we also introduce a second dynamic equation that describes the accumulation of production capital in region r , K_r , we get a simple model of multiregional development:

$$\dot{K}_r = s_r(1 - \tau_r)Q_r$$

where s_r signifies the investment coefficient in region r . We will not here discuss the equilibrium properties of this model. Instead, we limit our comments here to aspects of knowledge generation and knowledge flows.

We start with discussing the knowledge generation process in a region r , \dot{G}_r . It seems reasonable to assume that large and rich (and dense) regions will have advantages in terms of knowledge generation. They have a large output, can afford to devote large resources to R&D, have a larger capacity for learning-by-doing due to a larger output, and have normally larger knowledge accessibility due to a large intra-regional knowledge pool and a good accessibility to other large regions with a large knowledge pool.

What then can we say about knowledge flows and the potential for knowledge spillovers? It is obvious that the knowledge accessibility of a region is critical here. Regions with a high accessibility to knowledge-rich regions obviously have a higher potential for knowledge flows and pertinent knowledge spillovers. A high accessibility does not only imply a high probability for trade flows but also for direct investments and for interaction between people. These are all mechanisms that can generate knowledge spillovers and thus be a source for increasing returns in regional economic development.

22.5.3 Microeconomic Aspects of Knowledge Spillovers, Knowledge Externalities, and Regional Economic Development

Aggregated regional growth models offer general understanding of some of the basic factors that drive regional economic development. These models illustrate the importance of R&D, human capital investments, endogenous technological change, exports, agglomeration economies, and knowledge and other spillovers from a theoretical perspective, but they are too aggregated for empirical testing. This implies that these models do not have enough capacity to discriminate between different factors in terms of their importance for regional productivity and general regional economic growth. These models also have problems in identifying which the true exogenous factors are, that is, the factors that are of interest for regional policymaking.

Another problem with the aggregated models is that they assume that all economic agents are homogenous in all respects. Thus, these models fail to acknowledge that economic agents are heterogeneous in terms of their history,

age, size, knowledge and other resources, location, networks, ownership structure, routines, strategies, and behavior even if they belong to the same industry. Industrial and regional productivity growth and general economic growth are generated by a large number of different individual processes within the different economic agents, which are influenced by different internal and external. In every industry, there are economic agents that are persistently investing in R&D, while other economic agents invest now and then or not at all. The same is true for exports and imports. In terms of efficiency, we know that productivity distributions are characterized by a high degree of inertia and that the position that the different economic agents occupy in the distribution is highly persistent over time.

To overcome these deficiencies and to increase our understanding of how knowledge spillovers and knowledge externalities influence regional economic development, we need a model framework at the microeconomic level capable of accommodating different conditions and different behavior among economic agents within the same industry. Thus, we need a model framework capable of analyzing the factors influencing the behavior of different economic agents as well as the effects of differing behavior among economic agents in terms of productivity improvements and increased market shares. It is beyond the scope of this chapter to develop such a microeconomic framework with such a capability. However, in the sequel, we highlight some aspects that are critical for developing such a framework.

An externality in production emerges when the output from one economic agent is influenced not only by its own inputs but also from the outputs and inputs used by other economic agents. When the externality emerges within the region, where the economic agent is located, it has the character of a proximity externality. We talk about localization economies, if other economic agents in the same trade generate the externality. If economic agents in other trades generate the externality, we talk about urbanization economies. If the externality emerges from outside the region, we can talk about an extra-market or network externality. Assuming that economic agents are engaged in two types of production – the production of goods and services and the production of knowledge – we can formulate two functions that describe the two processes, the first being the production function for economic agent i , in industry j in region r :

$$q_{i,j,r} = A_{i,j,r}(c_{i,j,r}^q) f(x_{i,j,r}, y_{j,r}, z_{n,r}, E_s)$$

where q signifies output, $A_{i,j,r}$ the accessible knowledge for economic agent i , $c_{i,j,r}^q$ the absorptive capacity of economic agent i of production relevant knowledge, $x_{i,j,r}$ the inputs used by economic agent i , $y_{j,r}$ the accessibility to other economic agents in the region in the same trade as economic agent i , $z_{n,r}$ the accessibility to other economic agents in the region in other trade than i , E_s and the accessibility to economic agents in all other regions s . The above formulation implies that the productivity level in the economic agent i depends on its knowledge level, the size of the localization economies, the size of the urbanization economies, and the size of

the extra-market externalities. What about then the generation of knowledge by economic agent r ? Knowledge generation by economic agent i in trade j in region r is

$$\dot{g}_{i,j,r} = h \left\{ g_{i,j,r} \left[A_{i,j,r}, \dot{g}_{j,r}^A, \dot{g}_{n,r}^A, \dot{g}_s^A, \theta_{i,j,r}(q_{i,j,r}), c_{i,j,r}^g, \tau_{i,j,r}(q_{i,j,r}) \right] \right\}$$

where $g_{i,j,r}$ is the research productivity by economic agent i , $\dot{g}_{j,r}^A$ is the accessibility to knowledge output from other economic agents in trade j in region r , $\dot{g}_{n,r}^A$ is the accessibility to knowledge output in other trades in region r , \dot{g}_s^A is the accessibility to knowledge output in other regions, $\theta_{i,j,r}$ is the “learning-by-doing productivity” in economic agent i , $c_{i,j,r}^g$ is the absorptive capacity of economic agent i of R&D-relevant knowledge, and $\tau_{i,j,r}$ is the share of the output from economic agent i devoted to R&D. The current formulation tells us that an economic agent combines its current knowledge with new knowledge developed by economic agents in the same trade and in other trades in the home region, new knowledge developed in other regions, and with knowledge gained from the own production experience. The three knowledge accessibilities included represent three potentials for knowledge flows and thus for knowledge spillovers. However, the probability of each economic agent to succeed in the knowledge generation game is among other things dependent upon their skills and capacity in interacting with other economic agents, including, for example, research universities, in the home region as well as in other regions. It is also dependent upon the extent to which the location offers suitable interaction infrastructures and meeting places and formal and informal institutions including both a potential to protect the newly generated knowledge and the necessary level of trust for successful interaction between economic agents. The formulation allows for intra-market, quasi-market, and extra-market spillovers in line with the earlier discussion in this chapter. Concerning the extra-market spillovers, we want in particular to stress the role that exports and imports play as potential channels for knowledge spillovers. These spillovers can be direct, but they might also be indirect coming via other exporting or importing firms in the region.

The framework illustrated above is only a very partial framework. One important aspect missing is the creative destruction along Schumpeterian lines, that is, the scale and causes of entry and exit of economic agents and the role of knowledge spillovers in this respect. Innovation may represent a vehicle for new economic agents to enter the market successfully. However, since new firms by definition have done no R&D of their own, spillovers of scientific, technological, and entrepreneurial knowledge provide an explanation for innovative entry by new economic agents. Different regions have varying endowments of knowledge and not least new knowledge, which implies that the potential for innovative entry differs between regions and that larger regions generally offer a higher potential since they also offer a larger market potential. Of course, individuals and groups of individuals differ in their capacity to discover, create, and exploit innovations, that is, to create new combinations, but since individuals in larger regions generally have a higher education, more varied work experiences, and more extensive personal networks, these regions have an advantage here too.

22.6 Knowledge Externalities and Regional Economic Development: Policy Conclusions

Regional economic development seems increasingly to be dependent upon internal network structures to exploit fully the internal knowledge base, while at the same time absorbing new knowledge from outside the region. Given the potentially large importance of knowledge spillovers for regional (and thus national) economic development, policymakers have strong incentives to partly shift their traditional focus on the share of GRP and GDP invested in R&D to measures that foster knowledge spillovers, intra-regional as well as interregional. Concerning stimuli to intra-regional knowledge spillovers, policymakers should first of all focus on increasing intra-regional accessibility by means of investments in transport infrastructure and public transport. It is also important to establish arenas and meeting places where economic agents in the region can meet and interact.

Even if one should not underestimate the importance of intra-regional knowledge spillovers, it seems in particular important to promote interregional knowledge spillovers. There are certainly many channels for interregional knowledge flows. Starting with academic channels, we claim that it is important to involve scientists from other regions and countries in advanced research programs and to encourage scientists to engage in interregional and international cooperation and coauthorships as well as being guest researchers at research universities and laboratories in other countries. With a rapidly increasing number of patents and academic publications, it becomes more and more important to support inventors in all fields with rapid updates on new patents and publications. In particular, small- and medium-sized firms might need support when buying patents and licenses from other regions as well as in finding the right consultants in other regions. A positive and supportive attitude to strategic R&D cooperation with economic agents in other regions is also important. The role of imports of high-technology and knowledge-intensive products for interregional and international knowledge flows is often neglected. For regions that want to maximize the potential for interregional knowledge spillovers, it is important to create good conditions for import activities, not least through a high interregional accessibility. The same is true for direct investments from other regions. Multinational firms play an important but probably underestimated role for interregional knowledge spillovers. Another important channel for interregional knowledge flows is the in-migration of highly educated and skilled labor. Not least is it important to create an institutional framework that makes it easy for in-migrants to find information about available jobs, to apply for jobs, to find housing, etc.

22.7 Conclusions

There exists today a large literature on knowledge flows mainly using the concept knowledge spillovers without often making any distinction between intended and non-intended knowledge flows. The literature is not clear of the relative role of

intended, that is, market-based, versus non-intended, that is, spillover, knowledge flows. This distinction is important, since it is only the latter type which generates any knowledge externalities. The discussion of knowledge externalities in the literature has had a focus on technological spillovers, but it is obvious that there also exist pecuniary knowledge externalities. Thus, there is a substantial confusion in the literature as regards the economic nature of knowledge externalities. A similar confusion concerns the sources of knowledge spillovers. Many empirical studies have analyzed the role of patent citations, but it is obvious that economic agents learn from many other sources but patents. Another controversy concerns who benefits from knowledge spillovers. Is it mainly economic agents in the same trade or is it mainly economic agents in other trades? In terms of the mechanisms and channels for knowledge spillovers, it is obvious that learning takes place through many channels but that most studies only consider one channel at a time. A hot topic is the spatial reach of knowledge spillovers. It is a common statement in the literature that knowledge spillovers are bounded in geographical space, a statement that needs to be seriously questioned in the time of the Internet with global air connections, satellite TV channels, and a high volume of international trade, direct investments, traveling, and labor mobility. Furthermore, the consequences of knowledge spillovers and knowledge externalities are not well understood, disentangled, and analyzed. This state of the art certainly complicates analyses of the role of knowledge spillovers and knowledge externalities for regional economic development. A central problem here is that the majority of the studies are done at an aggregated level assuming that there is only one type of knowledge and that all economic agents are equal and benefitting to the same extent from knowledge spillovers. This implies that researchers in the field today have all too little understanding to guide policymakers. What we need for the future are multilevel studies based upon microeconomic data for economic agents, who recognize that economic agents and regions are heterogeneous and that different economic agents use different types of knowledge, use different knowledge channels, differ in their absorptive capacity, etc.

References

- Andersson ÅE, Mantsinen J (1980) Mobility of resources: accessibility of knowledge and economic growth. *Behaviour Sci* 25:353–366
- Arrow KJ (1962) Economic welfare and the allocation of resources for invention. In: Nelson R (ed) *The rate and direction of inventive activity*. Princeton University Press, Princeton, pp 619–622
- Audretsch DB, Feldman MP (1999) Innovation in cities: science-based diversity, specialization and localised competition. *Eur Econ Rev* 43:409–429
- Audretsch DB, Feldman MP (2004) Knowledge spillovers and the geography of innovation. In: Henderson JV, Thisse JF (eds) *Handbook of urban and regional economics*, vol 4. Elsevier, Amsterdam, pp 2713–2739
- Breschi S, Lissoni F (2001) Localized knowledge spillovers vs innovative milieux: knowledge tacitness reconsidered. *Papers Region Sci* 80:255–273
- Buchanan JM (1965) An economic theory of clubs. *Economica* 32:1–14

- Chambers RG (1988) Applied production analysis: a dual approach. Cambridge University Press, Cambridge
- Cohen WM, Levinthal DA (1989) Innovation and learning: the two faces of R&D. *Econ J* 99:569–596
- Cornes R, Sandler T (1986) The theory of externalities, public goods and club goods. Cambridge University Press, Cambridge
- Glaeser EL et al (1992) Growth of cities. *J Polit Econ* 100:1126–1152
- Griliches Z (1992) The search for R & D spillovers. *Scand J Econ* 94(Suppl):29–47
- Grossman GM, Helpman E (1991a) Quality ladders and product cycles. *Quart J Econ* 106:557–586
- Grossman GM, Helpman E (1991b) Innovation and growth in a global economy. The MIT Press, Cambridge, MA
- Helpman E (ed) (2008) Institutions and economic performance. Harvard University Press, Cambridge, MA
- Henderson JV (1974) The size and type of cities. *Am Econ Rev* 64:640–656
- Jacobs J (1969) The economy of cities. Random House, New York
- Jaffe A (1986) Technological opportunities and spillovers of R&D: evidence from Firm's patents, profits, and market value. *Am Econ Rev* 76:984–1001
- Jaffe A (1989) Real effects of academic research. *Am Econ Rev* 79:957–970
- Jaffe A, Trajtenberg M, Henderson R (1993) Geographic localization of knowledge spillovers as evidenced by patent citations. *Quart J Econ* 108:577–598
- Johansson B (2005) Parsing the menagerie of agglomeration and network externalities. In: Karlsson C, Johansson B, Stough RR (eds) Industrial clusters and inter-firm networks. Edward Elgar, Cheltenham, pp 107–147
- Kaldor N (1970) The case for regional policy. *Scott J Polit Econ* 17:337–348
- Karlsson C, Manduchi A (2001) Knowledge spillovers in a spatial context – a critical review and assessment. In: Fischer M, Frölich J (eds) Knowledge complexity and innovation systems. Springer, Heidelberg, pp 101–123
- Krugman PR (1991) Geography and trade. Leuven University Press, Leuven
- Lösch A (1954) The economics of location. Yale University Press, New Haven, CT
- Lucas R Jr (1988) On the mechanisms of economic development. *J Monetary Econ* 22:3–42
- Lundvall B-Å (ed) (1995) National systems of innovation – towards a theory of innovation and interactive learning. Biddle, London
- Marshall A (1920) Principles of economics, 8th edn. Macmillan, London
- Myrdal G (1957) Economic theory and underdeveloped regions. Duckworth, London
- Nelson RR, Winter S (1982) An evolutionary theory of economic change. Harvard University Press, Cambridge, MA
- Palander T (1935) Beiträge zur standorttheorie. Almqvist & Wiksell, Uppsala
- Rogers EM (1983) Diffusion of innovations, 3rd edn. The Free Press, New York
- Romer P (1986) Increasing returns and long run growth. *J Polit Econ* 94:1002–1037
- Romer P (1990) Endogenous technological change. *J Polit Econ* 98:71–102

Michaela Tripll and Edward M. Bergman

Contents

23.1	Introduction	440
23.2	Core Concepts	441
23.2.1	The Marshallian Concept	441
23.2.2	The Neo-Marshallian Industrial District Concept	443
23.2.3	The Innovative Milieu Concept	445
23.2.4	The Industrial Cluster Concept	447
23.3	Similarities and Dissimilarities	448
23.3.1	Geographies and Space	448
23.3.2	Actors and Interactions	450
23.3.3	Industries and Innovation	451
23.3.4	Environment, Competition, and Cooperation	452
23.4	Conclusions	454
	References	455

Abstract

Over the last three decades, literature on industrial districts, innovative milieus, and industrial clusters has enriched our knowledge about endogenous factors and processes driving regional development and the role of the region as an important level of economic coordination. This class of stylized development concepts has emerged since the 1970s and attempts to account for successful regional adaptations to changes in the global economic environment. Each of these

M. Tripll (✉)

Department of Human Geography, Lund University, Lund, Sweden

e-mail: michaela.trippl@keg.lu.se

E.M. Bergman

Institute for the Environment and Regional Development, Vienna University of Economics and Business, Vienna, Austria

e-mail: edward.bergman@gmail.com

concepts grew out of specific inquiries into the causes of economic success to be found in the midst of general decline by building upon the early ideas of Alfred Marshall in several ways. Neo-Marshallian districts found in Italy highlight the importance of small firms supported by strong family and local ties, while the innovative milieu concept places great emphasis on the network structure of institutions to diffuse externally sourced innovations to the local economy. Clusters have become far more general in scope, fruitful in theoretical insights, and robust in application, informing the work of both academics and policy-makers around the world.

23.1 Introduction

When one first becomes interested in the growth or development implications for local or regional collections of firms and industries, a bewildering array of possibilities presents itself. One learns that external economies of scale accrue to firms that collocate and thereby stimulate growth, for example, pecuniary savings and intangible flows of information arise automatically from agglomeration benefiting resident firms. External economies of scale are defined as economies that are external to the firm but internal to the region. External economies of scale can be divided into localization economies (benefits from colocation accruing to firms operating in the same industry) and urbanization economies (benefits from collocation accruing to firms operating in different industries). A robust and widely ranging literature of this tradition does exist, guiding the theoretically inclined or those who wish to examine empirical tests of important propositions.

Then there are the development practitioners, academics, or students who rely upon a family of more stylized models to formulate development prospects. Such models frequently adopt the concepts that are most widely accepted or formulated in their home territories, the bulk of which date from the 1970s. The emergence of these concepts can be documented by Google Ngrams for books published in English between 1800 and 2000 (<http://books.google.com/ngrams>). The 1970s was a period in which traditional production methods and their centers lagged but more peripheral areas, focused increasingly on better or higher technologies and market niches, began to prosper from the bottom-up. Italians are quite familiar with “neo-Marshallian industrial districts,” while the French and many Swiss prefer insights drawn from “innovative milieu,” and contemporary English or German speakers – and many others – are most comfortable with the ideas borne of “industrial clusters.” The concepts discussed in this chapter all go beyond a pure economic view on agglomerated industries (i.e., the argument of external economies of scale), drawing – although to varying degrees – attention to social and institutional factors that allow for coordination of economic actors. The preferred notions arose in specific contexts and circumstances, but are they essentially similar or are there elemental differences worth emphasizing?

In this chapter, we shall focus on both the common and distinctive elements of these concepts. To get under way, we shall first examine the core foundation on

which each of these rests, after which representative definitions will be drawn from the contemporary literature to distinguish between Marshallian and neo-Marshallian industrial districts, innovative milieux, and industrial clusters. We will then compare these with the definitions found in the emergence of their seminal literatures. Primary attention will be paid to features that reveal common overlaps and how each addresses these points, while also noting uniquely specific features that clearly distinguish among them. Finally, we will comment on challenges for future research and application.

23.2 Core Concepts

23.2.1 The Marshallian Concept

The English economist Alfred Marshall is seen as the “father” of the industrial district concept, and most contemporary work on agglomeration, localization, clustering, and the innovation-enhancing effects related with the geographical concentration of firms is explicitly or at least implicitly based on his writings. The key element in Marshall’s theorizing about “localized industries” (defined as industries concentrated in certain localities) is the notion of external economies of scale (see Sect. 23.1). It is these effects that enable small firms colocated with others in a district to compete successfully with large vertically integrated firms, which take advantage of internal economies of scale (i.e., benefits of large-scale production). There were ample examples of such geographical settings in the late 1800s in Britain such as cotton and textile in Lancashire, cutlery in Sheffield, pottery in Staffordshire, or straw plaiting in Bedfordshire.

Marshall’s writings contain both explanations for the rise of localized industries and their long-term “anchoring” in districts. According to Marshall, the initial localization of industries can have many sources such as the availability of raw material, the demand for goods of high quality or the immigration of people with specialized skills. However, once an industry is spatially concentrated in a particular locality, a set of agglomeration forces keeps it in place: “When an industry has thus chosen a locality for itself, it is likely to stay there long: so great are the advantages which people following the same skilled trade get from near neighbourhood to one another” (Marshall 1920, p. 225). These advantages are threefold, and they can be conceptualized as positive external economies of scale: (i) Knowledge spillovers: Firms benefit from local knowledge circulation and manifold opportunities for monitoring, learning from, and imitation of innovative actions set by colocated firms. Marshall (1920, p. 225) notes: “The mysteries of trade become no mysteries; but are as it were in the air, and children learn many of them unconsciously. Good work is rightly appreciated, inventions and improvements in machinery, in processes and in general organization of the business have their merits promptly discussed: if one man starts a new idea, it is taken up by others and combined with suggestions of their own; and thus it becomes the source of further new ideas.”

- (ii) Rise of supplier industries: Colocated firms operating in the same industry also take advantage from the growth of specialized supplier industries in their neighborhood, supplying the localized industry with raw materials, intermediate products, and services.
- (iii) Labor market effects: Localization of industries promotes the emergence of a highly specialized labor market that attracts both firms and workers. Employers benefit from the ready availability of highly skilled workers with the required special qualifications, and workers can take advantage of the rich employment opportunities, allowing them to find adequate jobs rather easily.

According to Marshall, the spatial concentration of firms of moderate size in districts was a widespread phenomenon – it was the typical form of industrial organization – of the English economy in the late nineteenth century, that is, a phenomenon that could be observed in many different industries and not merely a few.

Marshall's formulation of the industrial district concept contains a profound analysis of the linkages that knit collocated small firms together. His work points to knowledge spillovers, producer-supplier relations, and labor mobility among district firms and conceptualizes such relations as key factors underpinning the innovation capacity of localized industries. Marshall's work went beyond pure economic factors and considered also sociocultural and institutional assets to explain the economic vibrancy of industrial districts. A key notion in his writings is that of an “industrial atmosphere,” that is, the presence of a collective identity and shared industrial expertise that gradually develops in industrial districts and facilitates interaction, localized knowledge circulation, and the creation and diffusion of innovation.

Marshall was also aware of the potential dangers of specialized local development, anticipating the downturn of many English districts in the first decades of the twentieth century as well as what has happened many years later in old industrial areas characterized by economic mono-structures (i.e., overspecialization in mature sectors such as shipbuilding, coal and steel industries). He considered districts that are dependent on one industry as being extremely vulnerable, pointing to the risk of depression, crisis, and decline in case of changing context conditions such as, for instance, a fall in demand for its products or changes in technology.

In the era of mass production organized within large hierarchical firms, Marshall's ideas lost importance. His work on localized external economies of scale was brought to new life from the 1960s onward. Kenneth Arrow (1962) and Paul Romer (1986) extended Marshall's reasoning on knowledge spillovers as key source of innovation and growth. Such positive localization effects resulting from colocation of firms belonging to the same industry (intra-industry spillovers) are now known as Marshall-Arrow-Romer (MAR) externalities in regional science (Glaeser et al. 1992). They can and should be clearly distinguished from Jacobs externalities, that is, urbanization effects which arise from the presence of firms operating in different industries (interindustry spillovers).

The MAR literature provides a pure economic view of agglomeration phenomena, decoupled from Marshall's accounts of the supporting role of social and

cultural factors referred to as “industrial atmosphere” in his writings. However, Arrow and Romer were not the only scholars who have rediscovered Marshall. In the 1970s and early 1980s, Italian researchers started to draw extensively on his ideas, extending rather than downplaying the weight of the sociocultural underpinning of localized specialized economic activity in their conceptual and empirical analyses.

23.2.2 The Neo-Marshallian Industrial District Concept

The end of the golden age of mass production and the crisis of the large hierarchically organized firm in the 1970s provoked a renewed interest in Alfred Marshall’s work on external economies of scale and industrial districts. His basic ideas were revitalized and further developed by a group of Italian researchers who studied the industrial development in the central and northeastern part of Italy (the so-called Third Italy with its main regions Emilia-Romagna, Tuscany, and Veneto). This group was dealing with a somewhat puzzling phenomenon, that is, the economic success of industries with seemingly outdated organizational forms (family-owned small companies) belonging to mature sectors with limited growth prospects (such as textiles, footwear, leather goods, furniture).

Giacomo Becattini is one of the leading exponents of research on neo-Marshallian industrial districts. He rediscovered Marshall’s concept of industrial districts and reformulated it for the specific context of the Third Italy. Becattini’s article “From the industrial ‘sector’ to the industrial ‘district’” (published in Italian in the journal *Rivista di economia e politica industriale* (Becattini 1979) and in English 1989 in a book edited by Edward Goodman and Julia Bamford) is usually considered as the “starting point” of scholarly work on neo-Marshallian industrial districts (Landström 2005). The international dissemination of the concept was essentially promoted by Michael Piore and Charles Sabel. In their book *The Second Industrial Divide* (1984), they drew heavily on the example of neo-Marshallian industrial districts in the Third Italy to support their thesis of a major transformation from the Fordist mass production toward a model of flexible specialization. Looking at the Tuscan economy, Becattini (2003, p. 17) noted: “To employ a concept much used by Alfred Marshall, the course of Tuscan history leads to a form, still incomplete but already clear in outline, of “industrial district” . . . which produces economies external to the single firm and even to the industrial sector defined by technology, but internal to the “sectorial-social-territorial” network.”

Becattini and other protagonists of the neo-Marshallian industrial district notion (such as Gabi Dei Ottati, Sebastiano Brusco, Marco Bellandi, and Patrizio Bianci to name just a few) reject a purely economic view on local industrial growth and an exclusive focus on the economic effects of agglomeration. They suggest a much broader perspective that takes into consideration the social, cultural, and institutional foundations of local development. Becattini has advanced the idea of neo-Marshallian industrial districts as complex socioeconomic settings and highlighted the relation between efficiency and competitiveness of production and the

sociocultural conditions prevailing at the regional level. Becattini (1990, p. 36) defines neo-Marshallian industrial districts as “...a socio-territorial entity which is characterised by the active presence of both a community of people and a population of firms in one naturally and historically bounded area. In the district, unlike in other environments, such as manufacturing towns, community and firms tend to merge.”

Neo-Marshallian industrial districts in the Third Italy are characterized by a set of common features (see also Landström 2005). First (and accordingly emphasized in the definition presented above), in these districts, production activities and the “daily life” tend to merge. This merger is the outcome of a pattern of intensive interactions between the community of firms and families in a variety of dimensions. Second, there is a high division of labor among small companies, which are specialized in specific phases of the production process and work together in flexible teams. The manufacturer of the final good (the so-called impannatore) usually leads these teams and interacts with the market. Third, a widely shared value system (e.g., a specific ethic of work, principles of reciprocity) shapes action and interaction in the local community. These values are diffused throughout the district by a dense network of institutions (such as the firm, the family, the church, the local government, business associations). Fourth, in Third Italy’s neo-Marshallian industrial districts, competition and cooperation tend to coexist. The firms are linked to each other by manifold relationships, competing in some fields and collaborating in other ones. Finally, neo-Marshallian industrial districts benefit from a credit system with locally and socially embedded banks, which have a deep knowledge about and close linkages to the community of the districts’ firms.

From the 1980s onward, one could observe a rise in importance of local and regional policy interventions in Italian industrial districts. Policy-makers supported the evolution of districts by providing infrastructure (industrial parks, real service centers) and collective services (financing, education, marketing) and by performing the role of “social coordinators” (bringing actors together to solve common problems).

The 1990s has seen a growing skepticism about the long-term competitiveness and growth dynamics of neo-Marshallian industrial districts. Radical technological innovation and the globalization of the economy have led many authors to raise doubts about the extent to which they are a superior and stable spatial configuration of industrial production (see, for instance, Guerrieri et al. 2001). Critics have pointed to the limitations of self-sustaining production systems dominated by small firms in the global economy. Furthermore, the attempt made by Piore and Sabel (1984) to draw from very unique (and hardly generalizable) cases such as the Third Italy to develop their notion of flexible specialization has been viewed with extreme skepticism (see, for instance, Storper 1997).

Over the past years, a rich body of empirical work (for a review, see Rabellotti et al. 2009) emerged, documenting and analyzing ongoing structural changes in Italian districts. Based on a review of this literature, Rabellotti et al. (2009) found that some districts disappeared as a result of crises in their area of specialization (e.g., textile districts in Lombardy and Veneto). In most cases, these districts were

specialized in low-cost production and failed to compete successfully with manufacturers in newly emerging countries. Other districts have changed specialization, shifting from the production of final goods to the production of the machinery required for their manufacture. There are also districts which have conquered international luxury markets by upgrading the quality of their products, while others have enhanced their technological capabilities. Interestingly, the typical small Italian industrial district firm seems to lose importance. Indeed, the evidence suggests that it is medium-sized companies and groups of firms which are now the most dynamic agents and perform as key driving forces of structural changes (Rabellotti et al. 2009).

23.2.3 The Innovative Milieu Concept

The decline of many traditional industrial centers throughout Europe in the 1970s coupled with the subsequent rise of technological advances in industry and the emergence of innovative peripheral regions prompted serious rethinking of centrally directed and supported industrial development poles, often French in origin, in favor of innovatively driven indigenous growth. In the mid-1980s, Philippe Aydalot's hypothesized that there is "something" localized and intangible that permits innovation and dynamic development to proceed in certain regions and not in others (Crevoisier 2004). This observation set in motion a series of research efforts titled GREMI (Groupe de Recherche Européen sur les Milieux Innovateurs) to investigate and promote the dynamism seen in what became known as "innovative milieu." The concept of innovative milieu, however, has never ventured much beyond Francophone readers.

Research on milieus tends to be strongly focused on high-tech industries and on growth regions with extensive innovation intensity, although a few studies on conservative milieus in stagnating or declining traditional industrial regions and their restructuring processes also exist (see, for instance, Aydalot 1988; Maillat et al. 1997).

Definitions of innovative milieu vary, but most protagonists of this concept share the view that a milieu may be described as a set of region-specific rules, practices, and institutions that enhances the capacity of regional actors to innovate and to coordinate with other innovative actors (Storper 1997). In the writings of the GREMI group, the milieu is not always clearly distinguished from networks. As Storper (1997, p. 17) notes: "Many of the milieu theorists use the 'network' as their principal organizational metaphor. For some, the milieu is itself a network of actors ... in a region. For others, the network concerns the input-output system; it is the network that is embedded in a milieu, and the milieu provides members of the network what they need for coordination, adjustment and successful innovation."

The directions taken by GREMI members reflect what Benko and Desbiens (2004, p. 325) "see as a genuine 'territorial turn' that can be characterized by a movement from economics toward geography; this renewal is influenced by the cultural turn...that has been expanded through the influence of traditional

economics and sociology.” This reorientation has profound implications for how one looks at regional development. Accordingly, the focus moves away from MAR externalities enjoyed by individual economic units or agents and their industrial interactions to instead gravitate toward an examination of the full pattern of structural linkages among institutions that diffuse and enhance the innovative potential of a territory. Rather than product, supply, labor, or material flows among firms that constitute a local economy, one looks to the overarching structure of institutional networks through which innovations and ideas – often external in origin – pulse and bring prosperity to regions.

For example, Camagni (1991a, p. 3) sees an innovative milieu as “the set, or the complex network of mainly informal social relationships on a limited geographical area, often determining a specific external ‘image’ and a specific internal ‘representation’ and sense of belonging, which enhance the local innovative capability through synergetic and collective learning processes... The attraction of external energies and know-how is exactly the objective we assign to innovative networks: through formalized and selective linkages with the outside world...local firms may attract the complementary assets they need to proceed in the economic and technological race.”

This network is not seen as passive or merely contextual, but rather as a robust, proactive, and enabling agent-like presence: “...it is often *the local environment which is, in effect, the entrepreneur and innovator, rather than the firm*. The firm is not an isolated agent of innovation: it is *one element* within the local industrial milieu which supports it” (Aydalot and Keeble, 1988, p. 9). Despite considerable research from subsequent GREMI investigations, the question of how desirable externalities arose initially remains unclear; Simmie (2005, p. 793) concludes from his survey of the literature that “Explanations slip all too easily into the argument that innovative milieu assist innovative firms while at the same time the presence of innovative firms create the innovative milieu that are supposed to be assisting them.”

However one reads this literature, it would seem that firms and industries have become secondary or complementary components of a local economy, which rely upon the linked institutions of their innovative milieu to acquire key technological assets. Crevoisier and Maillat (1991) have traced some of the possible connections between the innovative milieu and the underlying markets, sectors, industrial organization, etc., giving rise to what has become more generally known as a *territorial production system*. However, the territorial production system (TPS) remains a largely unspecified “black box” of idealized categorical types (e.g., industrial organization), which does not clearly identify the agents, elements, or incentives that constitute a working local economy. Storper (1997) notes that one of the key shortcomings of the literature on innovative milieus is that it fails to identify the economic logic by which a milieu promotes regional innovation. Little is said about why localization and territorial specificity should promote technological dynamics.

Camagni (1991b) goes further in seeing the innovative milieu as an uncertainty-reducing mechanism that permits firms to better assess and deploy innovative

resources, although he does not refer to a TPS in specifying several gaps to be overcome. The innovative milieu concept first launched important ideas about how innovations are introduced and exploited in local economies, but it was eventually overtaken by other more convincing accounts of how innovation systems are populated and operate within regions.

23.2.4 The Industrial Cluster Concept

From the 1990s onward, the term “cluster” has become increasingly prevalent to denote spatial concentrations of firms. There are many different branches of this notion. As the term “cluster” is rather imprecise and meanings could differ, various authors have attached to it the adjectives “industrial,” “regional,” “business,” and “economic.” It is beyond the scope of this chapter to deal with all these variants (for an overview, see Iammarino and McCann (2006) and Cruz and Teixeira (2010)). In this chapter, we focus on Michael Porter’s cluster approach which has become the most popular one. In 1990, Porter published his highly influential book *“The Competitive Advantage of Nations,”* which introduced his *industrial* cluster concept broadly to business and policy officials, and he rapidly propelled awareness of the concept well beyond the previous academic audience in a subsequent outpouring of books and articles. Clusters are defined as “... geographic concentrations of interconnected companies, specialized suppliers, service providers, firms in related industries, and associated institutions (for example, universities, standards agencies, and trade associations) in particular fields that compete but also cooperate” (Porter 1998, p. 197). Porter claims that clusters are a dominant feature of the landscape, that is, they could be found in virtually all countries and industries. The concept is rather flexible as regards the types of firms involved. Some clusters are made up of small firms only, while other host both small and larger ones (Porter 1990).

Unlike the previous descriptive accounts of colocation from which causal factors might be inferred, Porter brings both normative and positive insights to his cluster definition from long experience as a theorist of corporate strategy at Harvard Business School. His cluster view essentially complements the importance of corporate strategies to remain prepared to dismantle and reinvest/relocate individual components of the corporate value chain (headquarters, distribution, production, etc.) in clusters of competitive and productive enterprises, which are comprised of similar or related firm components.

Porter’s main interest is in explaining the sources of enduring competitive advantages in a global economy. He argues that competitive advantage is strongly localized; it arises from clusters and is shaped by four determinants: (i) factor conditions (human capital, natural resources, infrastructure, etc.), (ii) demand conditions (sophisticated and demanding local customers), (iii) related and supporting industries (presence of capable suppliers and competitive related industries), and (iv) firm strategy, structure, and competition. These four determinants of competitiveness are interrelated and form a self-reinforcing system, the so-called diamond. Among all elements of the diamond, rivalry among cluster actors is seen

as the most important one because it has a stimulating effect on all others. Fierce rivalry promotes the development of highly specialized input factors, it upgrades domestic demand, and it spurs the rise of related and supporting industries (Porter 1998).

Vigorous competition among locally based rival firms is, thus, the main driving force in Porter's cluster model. However, he also states that clusters represent a combination of competition and cooperation, although the cooperative dimension of clusters is seen as less important compared to the competitive one. Porter's view on cooperation is rather specific. His writings reflect a deep skepticism when it comes to horizontal collaboration, pointing to a set of harmful effects that potentially arise from cooperation among competitors. Vertical cooperation (i.e., collaboration and knowledge exchange with suppliers, customers, and actors from related and supporting industries) in contrast is seen to be beneficial. Porter's work contains only some vague cues on the importance of social factors (more precisely, social capital) that facilitate interaction and collaboration between actors located in a cluster. Social (as well as cultural and institutional) features of clusters are clearly undertheorized in Porter's approach.

Other points of critique that have been raised (Martin and Sunley 2003) include – among other things – that Porter has failed to provide a clear specification of the geographical and industrial boundaries of clusters, related difficulties to identify clusters, and the uncritical reception of the approach in policy circles.

The undeniable attraction of Porter's cluster concept results from its twofold usefulness in providing answers to these questions: (i) What advice should a firm or corporate component consider in its corporate strategy to maintain maximum competitiveness when considering alternative business environments, and (ii) what advice should subnational regions consider when designing economic development policies aimed at providing attractive production sites and a sustainable employment base? Accordingly, many of the most ardent followers of this concept are business or economic development consultants and policy officials.

23.3 Similarities and Dissimilarities

It is already apparent from the brief outlines above that important differences but also similarities exist among the concepts considered in this chapter. To deal with these differences and overlaps more systematically and thereby reveal more about each, we will rely on a series of important features that are summarized in Table 23.1.

23.3.1 Geographies and Space

All concepts are based upon a specific understanding of space and certain geographic factors that help bound and define each in particular ways. The *Marshallian district* was essentially limited by distances that could be economically traversed

Table 23.1 Key features of concepts

<i>Geographies and space</i>	Marshallian industrial districts Geographical distance/ neighborhoods	Neo-Marshallian industrial districts Common political district occupancy with network	Innovative milieus Relational “space” coextensive with network	Industrial clusters MAR + globalized factors
<i>Core actors</i>	Small firms	Family/fluid firms, local officials	Social and political institutions, Firms of all size + industry groups knowledge producers	
<i>Agent motivation</i>	Profit maximization	“Mastery of craft” + place/product pride; consolidate high-value, luxury/income-elastic consumer markets	Growth, place/product pride; gain greater prominence in tech-based development initiatives	MAR + ROI, seize market and production opportunities afforded by globalization
<i>Interactions (relationships between actors)</i>	MAR	MAR, long-term, trust-based local networks	Informal social relations, collective learning within the milieu, formalized extra-local networks	MAR, supply/value chains, vertical forms of exchange, project alliances
<i>Industries, technological intensity of products, and markets</i>	All industries prevailing in the nineteenth century	Consumer discretionary/low- medium craft	High-tech intensive	Export-intensive industries/ high-tech and low-tech
<i>Innovation</i>	Diffusion of ideas, localized secrecies	Craft refinement and perpetuation, flexible organizations, incremental innovation	Adoption of extrinsically produced innovations that can be regionally exploited	MAR externalities, pre-commercial joint ventures, knowledge sourcing
<i>Environment (cultural and institutional context, extra-economic factors)</i>	“Industrial atmosphere”	Strong family and community values / identity, blurred boundaries between business and community spheres	Shared values and visions – regional cohesion	Trust and social capital forms the “glue” that binds cluster actors together
<i>Competition and cooperation</i>	Mix of competition (driving force of districts) and cooperation (precondition for collective innovation benefits)	Complex mix of competition and cooperation/high importance of cooperation as coordination mechanism	(Poorly specified) balance of competition and cooperation	Fierce competition as key driving force/cooperation less important

daily by workers and suppliers in the late nineteenth century, the densities of which were sufficiently high and distances sufficiently short that districts could often be referred to as “neighborhoods.” The late twentieth-century versions of *neo-Marshallian industrial districts* found in Italy are bounded instead within small governmental units known as political districts. The political boundaries are not matters of definitional convenience or regulatory requirement; rather, the local unit of government is an active partner and facilitator in the functioning of a neo-Marshallian industrial district. *Innovative milieus* are defined not at all by frictions of distance that separate the factors from individual producers or by official boundaries but instead by the “relational space” enclosed within a specific networked pattern of contacts among key institutions engaged in promulgating innovative impulses. *Industrial clusters* have perhaps the most flexible geographies, although they build upon Marshallian and MAR principles of market interaction that – with recent transportation and communication improvements – have expanded outward to include the “daily urban system” of entire metropolitan areas and regions. Indeed, Porter even speaks of a “California wine cluster,” although this is almost certainly comprised of several smaller, distinctive clusters of wine producers organized around a particular “terroir” that can take advantage of California’s full resources.

23.3.2 Actors and Interactions

Small- and medium-sized independent firms are the core agents in *Marshallian districts*. They are essentially profit-maximizing market actors who find greater market advantages by being located in industrial districts than if located apart. The prototypical *neo-Marshallian industrial district* found in the Third Italy consists of multiple agents: small family firms that expand or contract in various business arrangements with other firms, employees who may switch or hold jobs in multiple firms, and the local political district that establishes policies and practices. Profit maximization is but one of several motivations, which it shares with a mastery of craft, pride of place, and external recognition by consumers of luxury and income-elastic products. Agents are more difficult to isolate in the case of an *innovative milieu* because its constitutive networks engage many different institutions in a structure where specific roles go unremarked, although such institutions are said to be motivated by pride in their region and their efforts to promote innovation. Individual market and political agents operating lower down, at the level of the local economy, are seen as part of a “territorial production system” that interacts with and depends upon the innovative milieu, but these generally go unacknowledged as its core agents. *Industrial clusters* are essentially populated by firms and firm branches of all size and forms of corporate organization as their principal agents. It is often the case that the cluster itself acquires agent status and takes on a loose organizational form of institutional governance to promote the cluster, but this varies widely with fewer cluster organizations in Anglo-Saxon and more in continental European and Asian clusters. Return on investment, productivity growth, and market shares are principal agent motivations.

Looking at interactions, one finds that *Marshallian districts* are knit together by supplier linkages and often unconscious flows of knowledge, ideas, and workers (i.e., knowledge spillovers) among firms located nearby. The *neo-Marshallian industrial districts* theory puts due emphasis on long-term trust-based collaborative networks of small firms (and supporting organizations) that promote an easy exchange of tacit knowledge, joint purchase of materials, or joint initiatives to get access to technical or financial services. The literature on *innovative milieus* stresses the importance of mainly informal social relationships at the regional level which promote collective learning and, thus, the innovation capacity of companies. Collective learning is based on three mechanisms of knowledge transmission (see, for instance, Keeble 2000): (i) intra-regional mobility of labor, (ii) spin-offs, and (iii) formal and informal networks. It is noteworthy that the milieu literature specifically considers the role of extra-regional linkages. Such links to the outside world are more often than not formal in nature and considered as crucial in order to get access to complementary knowledge, markets, and technologies that are not available within the limited context of the milieu. The *industrial cluster concept* considers links of companies to demanding customers, suppliers, and vertical forms of exchange and interaction among companies. Both input-output relations and flows of knowledge are discussed in Porter's conceptual cluster model.

23.3.3 Industries and Innovation

Marshall did not specify the full range of 19th century industries, products, or technologies expected in his industrial districts, but it seems clear that any industrial process under way in England during that period could benefit from being located in a district. In short, the market interactions in *Marshallian districts* apply to all industries. Contrast this to the *neo-Marshallian industrial districts* of the Third Italy, which are very highly focused on mature industries and a more limited range of goods, markets, and production technologies. High-quality personal or household goods, often fashion- and design-intensive products, are intended for international markets of discerning consumers. Such goods are often craft-based or limited in production, which requires highly skilled artisans and quality-conscious production that takes advantage of incremental improvements in basic technologies. In comparison, the very idea of *innovative milieu* hinges on the defining importance of high technology, whose reemergence in the 1980s was seen as necessary to revive old (Maillat et al. 1997) and propel new production centers. Apart from the stress placed on links between high technology and industrial innovation, there appear to be no specific industries or markets implied. *Industrial clusters* are considered relevant to any industry or product where competitive forces require producers to enjoy MAR or Porterian diamond advantages. These are most acutely felt by firms competing in international markets, although locally competing firms are also said to benefit from cluster advantages.

In his analysis of British industrial districts in the nineteenth century, Alfred Marshall is remarkably clear about the sources and types of innovation arising in

highly localized industries. Innovation in *Marshallian districts* is seen to be essentially underpinned by a rapid diffusion of novel ideas and best practice solutions, spillovers of trade secracies, and a large stock of industry-specific knowledge which is constantly further improved by combinations of ideas of colocated firms and intergenerational knowledge transfer. Interestingly, Marshall seems to employ a rather broad definition of innovation that encompasses not only new products and processes but also “inventions and improvements . . . in general organization of the business . . .” (Marshall 1920, p. 225), that is, organizational innovations. Innovation in the *neo-Marshallian industrial districts* of the Third Italy is mainly incremental in nature (Asheim 2000). Small firm size, specialization in traditional industries, and little investment in formal research and development are seen as the main reasons for modest levels of technological innovation activity and radical innovation. Firms in Italy’s neo-Marshallian industrial districts tend to rely more on improving product quality, upgrading, and incremental process innovation (Rabellotti et al. 2009). Innovation activities are highly collaborative in nature (see Sect. 23.2) and typically oriented on craft refinement and perpetuation as well as on enhancing flexible organization structures. Protagonists of the *innovative milieu* concept stress both the importance of local collective learning processes (see Sect. 23.3.2) and the inflow of extra-local competences and complementary knowledge about technologies and markets. Innovation in milieus is essentially – although not exclusively – about the regional adoption and exploitation of knowledge, technologies, and innovations generated elsewhere. Innovation in *industrial clusters* is the outcome of the working of MAR externalities and conscious actions undertaken by firms such as pre-commercial joint ventures and acquisition of knowledge from a variety of sources. Elements of the diamond are seen to have substantial innovation-enhancing effects. Demanding customers, for instance, force cluster firms to carry out permanent improvements and innovation. Knowledge exchange and joint development projects with supplier and related industries provide essential inputs for innovative activities, and strong local competitors perform a critically important role in generating a high pressure to innovate. Taken together, these factors and conditions are supposed to constitute a fertile ground for continuous improvements and more radical innovations of products, processes, and organizations.

23.3.4 Environment, Competition, and Cooperation

The concepts under consideration here share an emphasis of the role that the environment, or more precisely, the cultural and institutional context, can play in regional development, competitiveness, and innovation. Marshall employs the notion of “industrial atmosphere” to highlight the role of social and cultural factors in supporting localized knowledge flows and industrial development. Colocation of similar firms in the same community implies that in *Marshallian industrial districts* “the secrets of industry are in the air.” A strong local cultural identity and shared industrial expertise are seen to form essential institutional foundations of the

evolution of Marshallian industrial districts. The literature on *neo-Marshallian industrial districts* clearly supports this view. Soft noneconomic factors such as a set of shared values, norms, trust, and collective identity are seen as fundamentally important to the economic success of industrial districts. The boundaries between the local community and industry are porous and often difficult to identify. Social and economic relations tend to be highly interwoven. Mutual trust, the identification with the region, and the products of a district, etc. are regarded as utmost significant. These factors enable and stabilize different types of collaboration and allow for a fruitful combination of cooperation and competition (see below). In a very similar vein, the *milieu approach* stresses the importance of the sociocultural dimension of innovation and development. A set of common values and shared visions (such as an orientation on long-term development goals instead of a short-term profit) and a willingness to collaborate reflect sound levels of regional cohesion and underpin high rates of innovation. It is important to note, however, that potential negative effects of too much cohesion on long-term innovation and adaption capacities of regions are also considered in the milieu literature. The *industrial cluster concept* proposed by Porter does not totally ignore noneconomic factors, but they are much less emphasized in his analysis when compared with the other approaches. He notes that trust and social capital are prerequisites for close interaction of cluster actors: “Social glue binds clusters together, contributing to the value creation process. Many of the competitive advantages of clusters depend on the free flow of information, the discovery of value-adding exchanges or transactions, the willingness to align agendas and to work across organizations, and strong motivation for improvement. Relationships, networks, and a sense of common interest undergird these circumstances. The social structure of clusters thus takes on central importance” (Porter 1998, p. 225).

One finds strong differences between the key concepts under review in this chapter when it comes to determining the relative importance of competition and cooperation as coordination mechanisms and key source of competitiveness. *Marshallian industrial districts* rely on both competition and cooperation. In Marshall's view, it is competition that is the key driving force of industrial districts, while district benefits in knowledge creation and innovation are a result of collaboration (Newland 2003). According to Marshall, cooperation can take two forms: it may be conscious and intentional (an example is the formation of industry associations) or unconscious and automatic (knowledge spillovers). In *neo-Marshallian industrial districts*, firms' willingness to cooperate is of critical importance for innovation and competitive advantage. In Third Italy's districts, “a complex mix of competition and cooperation” (Brusco 1990, p. 1) tends to prevail. “The efficient co-ordination of the district's activities and the promotion of dynamic growth is not simply a product of the unfettered operation of classic competitive market principles; on the contrary, what is at work is a complex amalgam of both competitive and co-operative principles . . . co-operation is at least as important as competition for organizing the district” (Sengenberger and Pyke 1992, p. 16). Protagonists of the concept of *innovative milieu* also emphasize the balance between competition and cooperation. The nature of this balance is, however, not clearly specified, and in

most contributions to the milieu school, the cooperative dimension seems to be rated as more important than the competitive one (Newlands 2003).

The *industrial cluster* concept – at least Michael Porter’s version of it – does not emphasize the role of cooperation to the same extent. On the contrary, it is fierce competition and not cooperation that determines competitive advantage. Porter and Ketels (2009) even argue that a large number of cluster benefits occur simply due to collocation, that is, for most benefits to unfold, no conscious and active collaboration by cluster firms is required. This does not mean, however, that cooperation is completely unimportant in Porter’s analysis. He recognizes some advantages in vertical forms of exchange, while cooperation between competitors is seen as harmful. “Clusters clearly represent a combination of competition and cooperation. Vigorous competition occurs in winning customers and retaining them . . . Yet cooperation must occur in a variety of areas . . . Much of it is vertical, involves related industries and is with local institutions. Competition and cooperation can coexist because they occur on different dimensions and between different players; cooperation in some dimensions aids successful competition in others” (Porter 1998, pp. 222). While cooperation is important, it does not have the same significance as competition. This argument is clearly supported by the fact that cooperation is not part of Porter’s diamond, while competition performs as the most crucial element in the diamond model because of the supposed stimulating effect it has on the other ones.

23.4 Conclusions

In this chapter, we have provided an overview on what is known about neo-Marshallian industrial districts, innovative milieus, and industrial clusters. From the 1970s onward, protagonists of these three concepts have essentially enhanced one’s understanding of the main sources of regional competitiveness and innovation. In the concluding section, we want to move beyond presenting past achievements made by adherents of the three concepts, glancing at the potential future of the notions under consideration. In our view, two aspects of future development deserve attention. The first concerns theoretical challenges; the second one is about the applicability of the concepts to other contexts and environments than those from which they have emerged.

A key issue for future theoretical work on neo-Marshallian industrial districts, innovative milieus, and industrial clusters will be to provide a more dynamic view on spatially concentrated industries and their institutional underpinnings. Although efforts are under way to understand their evolution, the literature on clusters and neo-Marshallian industrial districts in particular have been criticized sharply for relying too heavily upon static analysis and saying little about the development of regional collections of firms and industries over extended time periods. Consequently, a proper conceptualization of the evolution and transformation of agglomerated industries is urgently needed and should be a core topic of future research.

What is the potential future of the three concepts assessed by their transferability to other regions and countries than those from which they have emanated? Given the fact that the cluster notion provides a broader framework that allows for

capturing very different forms of geographical concentrations, one might argue that it will have the edge over the neo-Marshallian industrial district and milieu concepts. We see the cluster concept as the most fruitful one for cities and regions in Europe, North America, and Oceania and continuing in the near future, but it is unclear whether it will apply as convincingly in the more distant future or to other rapidly emerging economies of Latin America, Asia, or perhaps Africa. Is it a foregone conclusion that clusters as we now know them will arise in clearly identifiable forms in East Siberia, Mumbai, Jakarta, Amazonas, or Szechuan by 2050? Or will local circumstances – as was the case for neo-Marshallian industrial districts, innovative milieus, or industrial clusters – produce a rich variety of stylized models of local economic development to meet new challenges? Emergent countries may eventually come to differ rather considerably from the so-called Western models that presently operate within a relatively narrow range of characteristic democratic and market institutions. As the world's economies struggle to adjust to new monetary regimes, altered financial regulations, and impending resource frontiers, one cannot be certain how the systems of global production and international markets may change or how their knock-on effects may require the further repositioning of urban and regional economies.

References

- Arrow K (1962) The economic implications of learning by doing. *Rev Econ Stud* 29(3):155–173
- Asheim B (2000) Industrial districts: the contributions of marshall and beyond. In: Clark G, Feldman M, Gertler M (eds) *The Oxford handbook of economic geography*. Oxford University Press, Oxford, pp 413–431
- Aydalot P (1988) Technological trajectories and regional innovation in Europe. In: Aydalot P, Keeble D (eds) *High technology industry and innovative environments: the European experience*. Routledge, London, pp 22–47
- Aydalot P, Keeble D (1988) High-technology industry and innovative environments in Europe: an overview. In: Aydalot P, Keeble D (eds) *High technology industry and innovative environments: the European experience*. Routledge, London, pp 1–21
- Becattini G (1979) Dal settore industriale al distretto industriale. Alcune considerazioni sull' unità di indagine dell'economia industriale. *Riv Econ Polit Ind* 5(1):7–21 (Reprint of a new version in English: sectors and/or districts. Some remarks on the conceptual foundations of industrial development. In: Goodman E, Bamford J (eds) (1989) *Small firms and industrial districts in Italy*. Routledge, London, pp 123–135)
- Becattini G (1990) The Marshallian industrial district as a socio-economic notion. In: Pyke F, Becattini G, Sengenberger W (eds) *Industrial districts and inter-firm cooperation in Italy*. International Institute for Labour Studies, Geneva, pp 37–51
- Becattini G (2003) Industrial districts in the development of Tuscany. In: Becattini G, Bellandi M, DeiOttati G, Sforzi F (eds) *From industrial districts to local development. An itinerary of research*. Edward Elgar, Cheltenham, pp 11–28 (Reprint of Becattini, G (1978) The development of light industry in Tuscany: an interpretation. *Econ Notes* 2(3):107–123)
- Benko G, Desbiens C (2004) French economic geography: introduction to the special issue. *Econ Geogr* 80(4):323–327
- Brusco S (1990) The idea of the industrial district: its genesis. In: Pyke F, Becattini G, Sengenberger W (eds) *Industrial districts and inter-firm cooperation in Italy*. International Institute for Labour Studies, Geneva, pp 128–152

- Camagni R (1991a) Introduction: from the local ‘milieu’ to innovation through cooperation networks. In: Camagni R (ed) Innovative networks: spatial perspectives. Belhaven Press, London, pp 1–9
- Camagni R (1991b) Local ‘milieu’, uncertainty and innovation networks: towards a new dynamic theory of economic space. In: Camagni R (ed) Innovative networks: spatial perspectives. Belhaven Press, London, pp 121–144
- Crevoisier O (2004) The innovative milieus approach: toward a territorialized understanding of the economy? *Econ Geogr* 80(4):367–379
- Crevoisier O, Maillat D (1991) Milieu, industrial organization and territorial production system: towards a new theory of spatial development. In: Camagni R (ed) Innovation networks: spatial perspectives. Belhaven Press, London, pp 13–34
- Cruz SCS, Teixeira AAC (2010) The evolution of the cluster literature: shedding light on the regional studies-regional science debate. *Reg Stud* 44(9):1263–1288
- Glaeser EL, Kallal HD, Scheinkman JA, Schleifer A (1992) Growth in cities. *J Polit Econ* 35(6):1126–1152
- Guerrieri P, Iammarino S, Pietrobelli C (eds) (2001) The global challenge to industrial districts. Edward Elgar, Cheltenham
- Iammarino S, McCann P (2006) The structure and evolution of industrial clusters: transactions, technology and knowledge spillovers. *Res Policy* 35(7):1018–1036
- Keeble D (2000) Collective learning processes in European High-Technology Milieux. In: Keeble D, Wilkinson F (eds) High-technology clusters, networking and collective learning in Europe. Ashgate, Aldershot, pp 199–229
- Landström H (2005) Pioneers in entrepreneurship and small business research. Springer, New York
- Maillat D, Léchot G, Lecoq B, Pfister M (1997) Comparative analysis of the structural development of milieu: the watch industry in the Swiss and French Jura arc. In: Ratti R, Bramanti A, Gordon R (eds) The dynamics of innovative regions: the GREMI approach. Ashgate, Aldershot\Brookfield, pp 109–137
- Marshall A (1920) Principles of economics, 8th edn. Macmillan, London
- Martin R, Sunley P (2003) Deconstructing clusters: chaotic concept or policy panacea? *J Econ Geogr* 3(1):5–35
- Newland D (2003) Competition and cooperation in industrial clusters: the implications for public policy. *Eur Plan Stud* 11(5):521–532
- Piore M, Sabel C (1984) The second industrial divide. Basic Books, New York
- Porter ME (1990) The competitive advantage of nations. Harvard Business Review Press, Cambridge
- Porter ME (1998) On competition. Harvard Business School Press, Boston
- Porter M, Ketels C (2009) Clusters and industrial districts: common roots, different perspectives. In: Becattini G, Bellandi M, De Propriis L (eds) A handbook of industrial districts. Edward Elgar, Cheltenham, pp 172–183
- Rabellotti R, Carabelli A, Hirsch G (2009) Italian industrial districts on the move: where are they going? *Eur Plan Stud* 17(1):19–41
- Romer P (1986) Increasing returns and long-run growth. *J Polit Econ* 94(5):1002–1037
- Sengenberger W, Pyke F (1992) Industrial districts and local economic regeneration: research and policy issues. In: Pyke F, Sengenberger W (eds) Industrial districts and local economic regeneration. International Institute for Labour Studies, Geneva, pp 3–29
- Simmie J (2005) Innovation and space: a critical review of the literature. *Reg Stud* 39(6):789–804
- Storper M (1997) The regional world. Guilford, New York

Philip Cooke

Contents

24.1	Introduction	458
24.2	Regional Governance and Learning	459
24.3	Regional Governance and Policy Learning	460
24.4	The Learning Region	462
24.5	Regional Platforms, Methods, and New Innovation Policies	463
24.6	Regional Innovation Systems Version 3.0: Learning Dilemmas	465
24.7	Coevolution	465
24.8	Complexity	468
24.9	Emergence	469
24.10	Policy Emergence and Learning	470
24.11	Conclusions	472
	References	473

Abstract

In this chapter, an overview is presented of the three-phase evolution thus far of the regional systems of innovation perspective. The connected notion of the “learning region” is situated and subsequently re-situated in this account. The chapter begins by establishing the debate in the regional governance, learning, and policy contexts, especially with reference to the concept of “experimental regionalism.” Early reflections upon various critical responses to the 20-year literature on regional innovation represent the first main phase change, indicating the relative conceptual and empirical flexibility of the approach. Innovation in thinking about entrepreneurship is shown to have been at the heart of this first evolving perspective on regional dynamics. The most recent phase change represents the engagement of regional innovation systems, as a core subfield

P. Cooke

Centre for Advanced Studies, Cardiff University, Cardiff, UK

e-mail: cookepn@cf.ac.uk

of evolutionary economic geography, with key concepts in the complexity sciences. These are coevolution, complexity, and emergence. Each is shown to denote important new ways of thinking about regional innovation and evolution. The continuing relevance of the perspective to regional theoretical and policy application is underscored.

24.1 Introduction

There is now 20 years of solid theoretical and empirical research into regional innovation systems, and the concept is increasingly being applied in the world of policy analysis and practice. Regional innovation systems analysis has evolved through at least three versions (Cooke 1992, 2012; Braczyk et al. 1998; Asheim and Gertler 2005). The first phase change was from a Eurocentric, static, and manufacturing-led approach to a more flexible, dynamic and entrepreneurial approach. The second involved recognition of the importance of entrepreneurship in managing flows between knowledge exploration and exploitation (see below). The most recent phase change has been to fully recognize that regional innovation is an exemplary evolutionary process typical of complex adaptive systems as described by the likes of Kauffman (1993, 2000, 2008). This means a whole new vocabulary has to be comprehended that recognizes such processes as coevolution, self-organization, emergence, path dependence and path interdependence, relatedness, variety, and transversality. This is additional to but complementary with evolutionary economic geography terminology like related variety, search, selection and retention, mutation, speciation, and learning.

In what follows, an attempt is made to outline, critique, and elaborate key aspects of the above-mentioned phase changes in the evolution of a dynamic spatial research paradigm. It does this in a manner that intends to consider regional innovation systems in relation to the rather less-developed idea of “learning regions.” Both appear to have cognate origins, but a moment’s reflection shows the one to be proactive in its emphasis on innovation while the other looks reactive in its emphasis on learning somewhere else’s innovation. An effort is made in the chapter to reintegrate a more nuanced and advanced version of the learning concept. This is influenced by organizational practice based in complexity science. The classic regional innovation systems framework is constructed as follows. First, an open system architecture is proposed, which is the regional innovation system. Second, the system of innovation is composed of two subsystems: an *exploration* subsystem where research knowledge is both endogenously developed and imported and an *exploitation* subsystem where such knowledge is commercialized. Third, in- between is a “membrane” composed of intermediaries that may be “institutional” (mainly public, e.g., venture capital, incubators – expressed as an institutional regional innovation system or IRIS) or “entrepreneurial” (private services firms supplying such innovation support services – expressed as entrepreneurial regional innovation systems or ERIS). “Region” denotes the governance level between national and local. In any region, there is an assemblage of industries

that have distinctive technological trajectories and differential path dependences collectively referred to as the regional “paradigm.” Similarly, the region has an enveloping “regime” of hard and soft governance mechanisms influencing innovation as part of regional evolution.

In the section which follows, important aspects of the first phase change for understanding the regional paradigm or nexus of spatial economic processes are discussed. This heralded the emergence of the ERIS concept to balance the prevailing IRIS original. In the next section, some key implications of this for governance and learning at the regional regime or policy and regulatory level of analysis and activity are opened up. This moves the chapter into an assessment of the learning region notion, reasons for its apparent atrophy and a reassessment of its possible future role in a complexity-informed regional innovation systems (RIS) 3.0 model. Following this is a section that explores the foothills of Version 3.0 beginning with a critique of the industrial economists’ vertical, specializationist “framing” of economic processes. In this phase change the lateral concept of ‘platform’ is preferred to the vertical concepts of ‘sectors’ and ‘clusters’ for social agency involving innovation. Attached is an elaboration of a more appropriate, geographically informed ontology which is interested in horizontal interactions, knowledge recombination, and understanding innovation as involving “emergence” of novelty from unlike forms. This section, for which certain formulations are worked out for practice, is, like all good regional innovation systems research, theoretically informed and empirically tested with primary research data, but modeling is restricted to the conceptual level. This is because, on the one hand, modeling data are inappropriate in this context, but on the other, and more importantly, the complexity perspective, which actually derives from simulation modeling, has found that in the evolutionary sciences prediction – a prime justification for modeling socioeconomic systems – is impossible. Life itself, it concludes, is not subject to the predictive modeling achievements of physicochemical science precisely because humans are creative, innovative, social agents whose important future achievements cannot be predicted. Thus, evolutionary biological events can be understood *ex post* but not foreseen, except trivially, *ex ante*.

24.2 Regional Governance and Learning

Since the 1990s, a growth area in spatial analysis and practice has focused on regional innovation analysis and policy. To a remarkable extent, new problems and avenues for exploration emerge regularly regarding regional innovation processes and institutions, for example, intermediaries (Tödtling and Trippel 2011a, b; Nauwelaers 2011); variety, a key underpinning in respect of “relatedness”; and “conventions” – the soft institutions that inform culture and that are marked features of the new regional innovation challenge (Sunley 2011). These pose interesting tasks for modes of governance of regional innovation and demands for new kinds of learning, both more proactive than the “institutional borrowing” that characterized the supply-side era when markets became perceived as the solvents of

developmental dilemmas. Main results of the uncritical belief in the stability of markets in many countries have involved social polarization, financial market collapse, continued regional deindustrialization, if not industrial “desertification,” and dependence on now-eroding regional public sector employment to mitigate the resulting imbalances. Faced with the budgetary reckoning of this neoliberal experiment, regional governance, where it exists or survives, must persevere itself to become innovative.

Coordinated market economies (Austria, Germany, Nordic countries) have recognized this for some time, sustaining innovation support institutionally. The task was harder in liberal market economies, where injunctions that state intervention was the problem rather than the solution penetrated most deeply into the governance fabric. Heidenreich and Koschatzky (2011) reviewed the literature on regional governance of innovation, pointing to some fallacies and open questions about the manner of its conception and execution. These authors inhabit Germany’s coordinated market regime and are accordingly comfortable with federal norms that devolve some innovation and other knowledge responsibilities (e.g., universities) to the meso-level. They identify key efficiencies from knowledge recombination coordinated in regional institutions from the outset, primarily lower transaction costs, learning advantages from spatial proximity, and direct provision of “collective competition goods.” “Governance” moves beyond a region’s “soft institutions” such as conventions by addressing its “harder” government plus civic or associational governance regime. These can involve the nature of financial support for innovation (this can range from grants to loans); university coordination (e.g., regional mergers or centers of excellence); sectoral, cluster, or platform stimulus (see Harmaakorpi et al. 2011); training and skills formation; foreign direct investment; and regional promotion abroad, a nontrivial package of innovation instruments.

24.3 Regional Governance and Policy Learning

These authors, like many others, see building social capital as a target of regional governance. Variety in the interactions between paradigm and regime exerts a strong influence on the distinctiveness of regional governance idiosyncrasies, which extend to regional innovation system configurations. Although for complex systems to function effectively, there must be considerable system articulation, especially those involving multilevel regional-national-supranational strata; Heidenreich and Koschatzky (2011) also refer to studies that see considerable friction among such levels. This is caused by networks negotiating and bargaining about innovation according to distinctively layered democratic politics. Thus, although not hierarchically organized in a top-down manner, the supranational may still withhold resources from the national or regional levels if proposals to access policy funds infringe the “rules of the game” being targeted. Occasionally a region can reject national innovation policy inclination, clarify that it has reserved

powers, or move ahead with its own projects where the state has abdicated or allowed to fall into disarray, conceivably for ideological reasons, a national strategic responsibility. “Real service” provision to SMEs by Italian regional administrations was a case of the last named, which was subsequently forced into privatization by a hostile right-wing national government. “Regional experimentalism” after Sabel (2004) characterizes aspects of such friction resolution. In general, friction of the kind noted is a minority pastime.

Regional agglomeration and associated regional advantage arising from spatially proximate innovation, productivity, and growth also partly explain the success of ideas and practices promoting regional innovation governance. So does recognition by evolutionary social scientists and practitioners (for whom neoclassical theorems can seem otherworldly) of the difficult-to-measure value of social networking and “untraded interdependencies.” These constitute Storper’s (2009) regional “dark matter” measurable only by its gravitational influence. Enough is understood of these to at least see their effects in rapid regional mobilization that can swiftly translate identity into innovative action, showing a region has “got its act together.” Is this an immutable regional comparative advantage for some, or can it be “learned” for wider regional practice? The generic design for this is portrayed in Heidenreich and Koschatzky (2011) as dilemmas surrounding regional economic structure, regional networks, regional institutions, and regional policies. The “big shift” for new regional innovation policy is to attend to the content and multidimensional interrelationships of regional networks and institutions, in particular, rather than mapping structure directly onto policies and vice versa. Regional intelligence and policy learning thus suggest a more proactive, “catalytic,” or “orchestrating” role being required of regional innovation governance in future.

One crucial characteristic of the species evolution of regional innovation systems theory and empirics is that it has responded to the relatively few solid critical observations in an adaptable manner. Thus, as Tödtling and Trippel (2012) remind us, what we may call regional innovation systems version 1.0 (e.g., Braczyk et al. 1998) can, in hindsight, be seen to be somewhat Eurocentric in its emphasis upon public regional innovation intermediation and static in its portrayal of regional innovation system circuitry. These were products of the emergence of a new subfield (in both *regional* and *innovation* analyses) that began with European comparative regional research utilizing European-derived conceptual categories and generating tailored primary research data of a comparative kind. An important step forward was to recognize that other regional setups, though actually globally relatively few, were less “institutional” and more “entrepreneurial” in the provision of intermediary services (i.e., markets for innovation services were more developed, e.g., as in California or Massachusetts, Cooke 2007). To some extent, as also recognized by Tödtling and Trippel (2012), a dynamic element was introduced by returning for more longitudinal analysis 10 years later to re-research content for the primary regional innovation systems source book (Cooke et al. 2004). So this phase change we may refer to accordingly as RIS Version 2.0.

24.4 The Learning Region

One perspective that promised regional policy learning was “the learning region,” a concept developed by Florida (1995), adapted by Asheim (1996) and Morgan (1997) and recently reviewed by James Simmie (2011). Somewhat disappointingly the promise of this notion has not materialized, partly because as Simmie shows it got bifurcated into a normative idea, resting on the injunction that learning was a desirable end for regions to aspire to, on the one hand, and more empirically that it was a modest action line in regional innovation *strategy*, on the other. Accordingly, it has never developed analytically even though, as much of the regional innovation systems literature makes clear, there is evidently an acutely perceived need for better qualities of “learning” by firms and other innovation actors and for “regional policy learning” to tackle issues such as “cognitive dissonance” among corporate functionaries and entrepreneurs, “convention analysis” of regional production culture, policy mixing to stimulate “path creation,” and the hybrid skills to facilitate relatedness and transversality among policy functionaries. It will be shown below how in the evolution to RIS Version 3.0, learning techniques and instruments have been refined to facilitate it in both RIS and learning contexts.

Correctly, in its origins, in the work of Richard Florida in 1995, “learning region” is a response to the rise of the knowledge economy as is even more the case in a rapid follow-up article by Björn Asheim which grounds the notion in Lundvall and Johnson’s advocacy of building a “learning economy” to face the exigencies of the same phenomenon. Michael Storper did not write about “learning regions” as such but devoted research time to comparing “technological learning” in clusters with different convention sets or modes of untraded interdependence, which were probably the most fruitful theory and practice lines to follow. Finally, probably the most-cited variant of the “learning region” idea was Kevin Morgan’s paper of 1997. Here, Simmie shows the key to regional regeneration and improved social welfare lay in strengthening a region’s social capital and institutional capacity to support learning.

Critique of the concept has ranged from ascribing it the status of “fuzzy,” an “impressionistic neologism,” “unlikely,” “over-localized,” and challenged by “learning asymmetries” (see, e.g., Martin 2001). This is reminiscent of the many critiques aimed at the cognate concept of “organizational learning” in the large corporation. Here, problems concerning how to sample, from where, or from whom to learn, whether what might be learned was applicable, and indeed whether it was yesterday’s knowledge, meaning the learner would be engaged in a permanent failure to “catch up” were all raised. This all seems rather unfair to what – if the concept had been better specified, perhaps in terms of learning the region’s paradigm and regime uniqueness and how it might be “nudged” toward path interdependence – we now see to be a fundamental cognitive need in accomplishing regional innovation and growth. This seems to be the thinking in recent attempts to revisit the concept by Rutten and Boekema (2007). However, the kind of conceptual and policy instruments needed to achieve such endogenous regional change remain to be clearly specified. We shall see below how some progress in this regard has

been made from the viewpoint of complexity theory (Mitleton-Kelly 2006; 2011). One task of this chapter is to evaluate such progress from the viewpoint of RIS Version 3.0.

24.5 Regional Platforms, Methods, and New Innovation Policies

This task was begun in the early 2000s, culminating in Harmaakorpi et al. (2011) with their “platform” concept of regional innovation and renewal. The roots of this model are found in recognition of basic Schumpeterian insights into the nature of innovation as a product of cross-fertilization (recombination) of knowledge and ideas. This is something which the cluster idea, as the apotheosis of proximate specialization, obscured for academics and policy-makers alike for two decades or more. Accordingly, there has been a lack of policy measures to foster practice-based, networked innovation processes that combine diverse knowledge bases. It could be added that until recently, and in the process of being articulated into a synthesis here, there has been relatively little intellectual leadership of an alternative perspective either. One reason for this is Lundvall’s line that policy learning is for policy-makers not academics, which on a moment’s reflection is a little timid. For while it is always difficult to think out and design policy abstractly in the “ivory tower,” a thoroughgoing critique of the economist’s traditional vertical “framing” of spatial processes opens up significant opportunity for innovative policy guidance. Harmaakorpi and associates are by no means alone either in having the privilege of occupying both “worlds”: academia and innovation policy-making simultaneously. This enabled them to conduct “regional experiments” à la Charles Sabel. Out of this experience, a newly minted criticism they discovered of the “proximity” perspective is that it fails to explain how learning from knowledge spillovers actually happens and that the effect of these may be negative. They find the distance implied in the notion of variety more appealing because it avoids negatives like involuntary spillovers, opportunism, and lock-in. Accordingly, their aim is to create an efficient *balance* between the contradictory purposes of enhancing proximity and distance. The cluster model is seen to be suboptimal in this respect and accordingly inferior to a platform model of regional innovation policy.

The “platform model of regional innovation policy” displays the following key characteristics: Its network morphology is one based on *loose coupling* of weak ties engaging with “structural holes.” Structural holes are the spaces (sometimes “white spaces”) between industries or specialized clusters. As we shall see later, loose coupling is an essential property of innovation-inducing adaptive systems from a complexity theory perspective (Mitleton-Kelly 2011). These are areas where network interactions may produce innovations if the holes can be bridged with innovation discourse, action, and content; social capital is thus of the institutional “bridging” kind; knowledge production is transversal; knowledge conversion is by means of cross-fertilization; regional absorptive capacity is future-oriented; experience-based learning is favored over science-based learning; external economies

are those of “urbanization” rather than “localization” in kind; and innovation systems are regional and national. On this basis, structured experiences of challenges and change of conventions, competences, and capabilities are induced by articulation of discourses among firms and stakeholders combining related knowledge from inside and beyond the region. The aim of this rendering of RIS 3.0 is to create a regional platform based on relatedness and supported by platform policies that optimize it to optimize innovation.

Facilitation of the required articulation of discourses that may valorize or change conventions and build up firm and intermediary competences and capabilities is intended to instigate a structure of learning institutions and processes in the region. The key spatial process aims here are as follows: to clarify the nature and forms of regional-related variety; to facilitate the *recombination* of knowledge; to identify the “structural holes” or “white spaces” where innovation opportunity may lie; and to evolve regional platforms that combine knowledges, clusters, and sectors for purposes of innovation. Complementing these in a new regional innovation systems perspective are four policy concepts. First, “enlightenment” may be diffused through the deployment of dramaturgy, literally acting out scripts of representative “convention sets” under challenge; second, assistance comes from having a mode of “orchestrating” dramaturgy and other learning facilities, such as “ideas incubators,” “living laboratories,” and “improvisation sessions.” Orchestration here implies promotion of such assets and conducting their articulation into a coherent narrative. Third, innovation system integration comes from achieving “transversality” or the cross-pollination of intercluster or sector-cluster innovation potentials within and beyond the region, and finally, of key importance is evolving methodologies, such as technology or creativity matrices to concretize commercializable innovation actions and outcomes. The exemplar of this is engineering-rich Bavaria, but it is also practiced in the design-driven innovation context of Lombardy and its creative and innovative design-intensive domestic furnishing, lighting, and kitchenware clusters as described by Verganti (2006). Here, the innovation paradigm is changed relatively frequently and radically in the “episodic” sense by articulation of discourse that changes conventions through changing the meanings prioritized in the prevailing “sociocultural regime.” This demands inputs both from expert “circles” inside and well beyond the region and within and well beyond the specific cluster. It requires strong articulation of regional firms and stakeholder institutions, and it is “orchestrated” in ways that “propose” innovations to markets. It can thus be vulnerable to overestimation of the market appeal of new lines, but such “practical reasoning” is also built into the articulation of discourse process.

Accordingly, a new paradigm for regional innovation and growth has been evolved in ways that meet the criticisms of the weaknesses and *lacunae* of inherited models, rooted as these initially were in Eurocentric regions, statically described and under the influence of manufacturing supply chain thinking of the 1990s. This is by no means the only way forward, but it resonates completely with the main threads of the discourse from its Schumpeterian origins to the modern day. Accordingly, the new agenda for regional innovation policy is different from the old.

As others have noted, endogenous innovation policy with the regional agency in a more catalytic role is now expected to replace the backstop functions of old. These had evolved in responding to market failures and welfare enhancement imperatives in the neoliberal, supply-side era. This is conceived as the appropriate posture in the context of a global knowledge economy regime assailed by seemingly intractable crises of economy and ecology. The region where innovation platform methodology was pioneered was a declining economy within Finland; the region where pulp and paper relatedness evolved to cross-media clustering is a relatively poor, peripheral region in Sweden; London and other metropoles may not be as innovative as presumed because they bask in conventions of entitlement, expectation, and privilege. Accordingly, the traditionally conceived “innovation paradox” in which least absorptive capacity is found in regions needing most innovation is clearly somewhat shaky and in need of measured reflection. This draws attention to the problematic nature of “smart specialization” as a “learning” response. Rather than implementing inappropriate initiatives from a supranational hierarchy, regional systems grow best by “emergence” of innovation from the recombination of their own paradigm and regime assets.

24.6 Regional Innovation Systems Version 3.0: Learning Dilemmas

One of the strong criticisms of the advocacy for “learning regions” a decade or more ago is that they were implicitly or explicitly modeled on an exemplar, usually Silicon Valley. The reason why this “framing” of the problem always produces disappointing results is aptly summed up below:

...The [organisational re-design] process was systemic and could not be reduced to individual parts or components or specific individuals on their own. That is part of the reason why “best practice” cannot be copied. The process is systemic, emergent and context dependent. It cannot be reduced to “building blocks” which can simply be re-assembled in a different context and give rise to an identical outcome. ... (Mitleton-Kelly 2011, 49)

This criticism is expressed from a complexity science perspective, which is wholly compatible with the kind of regional innovation system analysis and practice described in the preceding section. In order to explore the new take on regional innovation systems further, the following sections will explore the analytical and policy relevance of the approach by focusing on three of its master concepts: coevolution, complexity, and emergence. The revitalized role of “learning” in regional development is considered in the section on complex policy learning.

24.7 Coevolution

A good example of a coevolutionary analysis is Murmann (2003) who compares the evolution and institutional interactions involved in the separate fates of the British,

German, and American chemical industries. He finds that the coevolution of science and industry in Germany was a crucial coevolutionary series of events in that industry's success. However, in recent years, coevolutionary thinking came to the fore in the eco-innovation field. Here the intellectual effort was devoted to trying to understand and facilitate the transition of global society from its traditionally carbonized energy systems toward a non-carbonized, renewable energy future. Of special spatial interest has been why rising global concern with climate change issues produces national and regional policy responses that vary from the concerned and enthusiastic to the apparently unconcerned and apathetic. The idea of coevolution, and its absence is germane to such variable outcomes, a conclusion of Unruh (2000) who described a politico-economic institutional regime that has produced worldwide "carbon lock-in" as much for institutional as economic reasons. The US sub-prime financial crisis demonstrates negative coevolution (or systemic positive feedback) perfectly. Accordingly, the political subsystem, the consumer subsystem, the construction subsystem, the financial subsystem, and the energy subsystem were all coevolving in a particular, "dominant design" modality. Elsewhere, things were different, and a region like North Jutland and its country, Denmark, had simultaneously begun to express their "emergence" away from "carbon lock-in" through eco-innovation, initiated in the regional paradigm, whereby agro-engineering capabilities (milk cookers and turbofans) "emerged" into wind turbines, giving the innovating region the status of a "transition region." This meant it could innovate away from "carbon lock-in" by recombining a well-developed regional eco-innovation paradigm. Being home to former agro-engineering firm, *Vestas*, the world's leading wind turbine producer, *Grundfos*, a leading photovoltaics (solar energy) exporter, *Velux* (insulated windows), its owner *Arcon* (a leader in biogas energy production), and numerous green engineering SMEs allowed dynamic knowledge capabilities to be recombined in sustainable combined heat and power (CHP) design and construction. This capability was embedded in a pervasive "green" sociocultural and consumption regime supportive of local renewable energy networks (Cooke 2010). One complementary way to understand this process is according to a coevolutionary transition model (Geels 2006).

Coevolutionary transition theory, even with its multilevel perspective (MLP) is intellectually interesting on the one hand but frustrating on the other. The intellectually interesting level concerned is the process whereby globally significant innovation rose to prominence if not yet dominance at the level of the socio-technical system (STS). Theory suggests such innovation is destined eventually to become the dominant design (as hydro, solar, or wind power are for renewable energy today) and ultimately take over from carbon. Evolutionary economic geography theory also opened up this coevolutionary vein of research as evidenced by Martin and Sunley (2010a) who had critiqued classic path dependence theory as static and equilibrium orientated, opening up the prospect of a more dynamic perspective on regional development based on path *inter*dependence. However, their approach lacked a convincing mechanism for bringing such novel states about. Similarly, the frustrating aspect of the earlier STS approach to transition was that it lacked a causal mechanism, change being seen as unproblematically arising from

market transactions or something akin to “enlightenment.” Reflecting upon this for path interdependence, it seemed primarily because, like much evolutionary economic geography, the nevertheless interesting and creative insight lacked a convincing theory of *innovation* as distinct from a vague notion of “technological change” as being somehow involved.

The key thing about a complexity analysis (which the coevolutionary transition model is not) is captured in the following observation by Eve Mitleton-Kelly that is relevant also to the broader project of evolving a richer theory of regional innovation and development:

...The distinguishing characteristic of complex co-evolving systems is their ability to create new order. In human systems this may take the form of new ways of working or relating, new ideas for products, procedures, artefacts, or even the creation of a different culture or a new organizational form. ... (Mitleton-Kelly 2006)

The way forward here is helped by “reframing” the theoretical problem as a transition from thinking of path dependence to one of conceptualizing path interdependence. This is integrated to another core concept in evolutionary economic geography (EEG), namely, “related variety” (Boschma and Frenken 2003; Frenken et al. 2007). These authors showed empirically that regions with industries in neighboring sectors (North Jutland’s eco-industries would be an excellent illustration) benefitted from a double “proximity effect.” The first of these is a *relational* advantage, which facilitates exploitation of “knowledge spillovers” because of the high lateral absorptive capacity potential of firms toward each other’s external economies of information. The second effect is in terms of the *geographical* proximity that facilitates by time-space compression the aforementioned *relational* advantage. This enriches information such that its elements of difference and surprise (“news”) may be communicated and factored into innovation calculations early, even before their full meaning has had time to be realized commercially.

This idea about the nature of information in innovation makes a significant contribution to RIS Version 3.0. It explains how coevolution of path-dependent processes can combine in order to branch into new path creation through facilitating path interdependence. The small but crucial addition that has to be made, from a spatial perspective, is that even though the relevant message may come from a great distance *geographically* or *relationally*, it has to be exploited in a particular space or place – the location of the innovation *design*. Such a location may take the form of a “transition region” as discussed above. Many innovations display this characteristic of combining or recombining information from widely different sources in a place that is nevertheless nonrandomly “selected” and explicable in terms of path dependence and path intersection of STSs. One of the key contributions the perspective makes is to expand the meaning of “related variety” beyond the narrow confines of neighboring industries such as electrics and electronics, automotive and aerospace engineering, or banking and insurance. This means speaking of “relatedness” more generally, encompassing both routine and possibly surreal knowledge combinations for specific innovation. Information, even devoid of semantically precise meaning, is capable of making a *difference*. This means that

the unexpected interest or surprise that even *information* may provoke may help solve a problem related to the tendency to disorder (entropy) faced by the social agent seeking knowledge to innovate. The strong element of *surprise* involved here means that innovation prediction is impossible except in relatively trivial ways. Accordingly, “related variety” effects may be hypothesized *ex ante*, but they may only satisfactorily be understood *ex post*. This is called “revealed related variety” and captures the strong element of unexpectedness and unpredictability that seems to be associated with most innovation. This occurs with increasing frequency due to the expansion of “cocreated” variety in economic evolution, which means novelty becomes both more widespread and easier (Kauffman 2008, 151–154).

24.8 Complexity

There are clear resonances between the coevolutionary perspective that also incorporates key concepts like path dependence, related variety, and relatedness from EEG and the key findings of the complexity sciences (see, for an early economic geography approach to complexity, Martin and Sunley 2010b). One key difference between that treatment of the spatiality of complexity science and the present one is that this one relies significantly on complexity theory with an evolutionary biology inflection while, the other is informed by more of a physicochemical systems model. This is important because, as noted, Kauffman (2008) shows that evolutionary biological processes like selection, speciation, and mutation are unpredictable. By contrast, planetary and subatomic movements are largely predictable, albeit surprisingly often vitiated by data difficulties and even cavalier attitudes by scientists toward data where they do not fit the mathematics.

A second area of agreement between coevolutionary and complexity theory concerns the element of *difference* referred to above as being of such importance. This applies even in the analysis of the manner in which “mere” information, let alone meaningful *knowledge*, contributes to cognitive combinations and recombinations. The complexity science explanation of path interdependence is conceived of as occupying an imaginary topological landscape characterized by “strange attractors.” This is because complex adaptive systems are conceived to have an “ontogenetic” topology or “fitness landscape.” This fitness landscape can be rugged or sleek and variations in between. The sleeker the landscape, the more stable the system because there are few sources of perturbation and little opportunity for communication between system entities. This epitomizes the “wilderness” region with few sources of economic energy with which to interact. The more rugged the landscape, with metaphorical valleys and their tributaries acting as communication lines between centers of energy or potentially interacting entities, the more potentially unstable is the system. This is in the sense that it is prone to disequilibrating “collisions” of economic activities or their sub-elements that give rise to novelty. Some such interactions are considered to occur between “normal” attractors (or “routine”-related variety from an EEG viewpoint), but others involve “strange attractors” that are unexpected or surreal combinations that nevertheless

find ways to combine or recombine into innovative pathways. In Kauffman (1995), such centers of energy in complex adaptive systems are called “clusters,” and while these are different from the usage in economic geography, the latter are nevertheless a good illustration of the former. When such interactions are abundant, the system is said to be operating at “the edge of chaos.” This does not mean it is an utterly disorderly space but rather a condition in which the kind of system change, novelty, or innovation called for by Mitleton-Kelly (above) can occur. Finally, the complexity perspective also helps open the black box of innovation because of two core concepts introduced by Kauffman (2008), namely, “preadaptation” and “the adjacent possible.” These are options from within the complex adaptive system’s “normal” or “strange” attractor subsystem elements or “clusters” that are seeking novelty from the interactions that “the edge of chaos” has provided. In Kauffman (1995), he talks about these, naturally enough, although in complexity science it is, to be sure, a rare enough occurrence, in terms of communication between persons. Thus, interactions may initially take on the informal status of “gossip” between even lower-order employees of two incumbents (firms) in different entities (clusters or industries). Connectivity of this kind may reach middle managers in the strange attractor companies who might be *surprised* at the information passed on to them for semantic interpretation with senior executives interested in solving an innovation problem.

One direction such deliberations might take involves “preadaptation” whereby an innovative practice, product, or service implemented or marketed by a firm in one distinctive industry might already have within it sufficient information to allow it to be reworked in the other distinctive industry or cluster. This is both a not uncommon way in which usually incremental innovation actually occurs, and a strategy of how some “ahead of the curve” intermediary, innovation support agencies define their function in the regional innovation system. A good deal of such effort can involve “creative” activities like “sensemaking” of the kind Weick (1995) writes about, to “storytelling,” and even “critical theater” after Schreyogg and Hopfl (2004). This may seem strange to audiences unfamiliar with corporate change management practices or those of regional innovation agencies in countries that habitually make use of living laboratory learning and training settings. The second direction the innovative mind has the opportunity to follow is described as “the adjacent possible” where a step or steps into the unknown seeks to bridge the gaps where innovation potential might lie. This may lead to radical innovation where many sub-innovations may spin off an initial breakthrough, or it may be disruptive where some change in product status is induced in the appropriate market (online financial services, budget airlines, etc.), or it may be incremental but nevertheless an improvement to current practice. Evidence of both kinds of strategic innovation advice and practice are presented in the final brief section of this chapter that precedes the conclusions.

24.9 Emergence

This is a cognate concept to coevolution and complexity that provides theoretical interest but also gains additional practical meaning from its engagement with

regional innovation systems and practices. It has also usefully been reviewed by Martin and Sunley (2012) albeit from a fairly conventional top-down perspective. In “emergence” theory, the higher level tends to have been seen as the one responsible for qualitative change in elements that already exist in independent form at the lower levels of magnitude. But from an evolutionary biology perspective, the lower levels are usually determinant. Rarer is evidence of top-down causality. The examples of sugar or water existing at a superior level to that of the molecules that comprise them are often utilized as an illustration of emergence in the physicochemical world. The key point, however, is that “emergence” is caused by transversality rather than simple additivity. Transversality unites horizontally the properties latent in “relatedness” of natural or strange attractors. Thus, exploiting *difference* is actually at the heart of both “innovation” and “emergence”: Indeed they may, from a regional innovation systems perspective, be interchangeable.

In the economic geography literature, the question of “emergence” has been directed at, for example, the issue of cluster emergence (Fornahl et al. 2010). Hence, we might want to explain a cluster’s existence in terms of its agglomerative scale, which is a quantitative matter, but in terms of “emergence” the phenomenon under inspection is not scale dependent but relational. If colocated firms in the same field are working together on a regular basis, they can be a cluster. Accordingly, it is then a question of finding out why they find collaboration, colocation, and cocreation agreeable business strategies rather than how they simply came to agglomerate in space. The latter is an interesting question about *agglomeration* (which typically lacks collaboration and cocreation), but not especially about clustering. In other words, the cluster is “emergent” from the shared interests of the elements in higher-order economic activity; they could not achieve acting alone just as sugar is formed from but more than its constituent, lower-order molecules. Accordingly, it is as clear that the cluster elements collocate in space as the necessary sugar molecules do. The key point for regional innovation systems here is that when not interacting to create sugar, carbon atoms are available to bond with hydrogen atoms to make water or innumerable other chemical compounds used in everyday life. In other words, their “existence space” is the basis upon which their innovative recombination operates. Equally, some such atoms (or firms) may like to collocate, but not cocreate.

24.10 Policy Emergence and Learning

The exposition of RIS phase changes given above invites questions regarding the validity of its key propositions about coevolution, complexity, and emergence. A research project was, accordingly, implemented in Sweden, where complexity theory-derived measures were being deployed in two out of three regions studied. Thus, the research material alluded to in this section on policy was elicited from face-to-face interviews conducted with three regional development agency heads and some 12 cluster intermediaries in three Swedish regions during early 2011 (Cooke and Eriksson 2011). Briefly, the following case comparisons show instances both of “emergent” policies interlinking different activities at local level into

a grander synthesis at regional and even national levels. They also reveal, in one case, policy “learning” that leads to a complexity variant of “bifurcation” toward “clusters” of energy in a region that show more economic potential than the formerly path-dependent trajectory. Hence, the “emergence” perspective, informed by coevolution and complexity, begs some questions we hope to answer. At which system level does initial causality lie when, for example, the phenomenon under inspection is policy agency to seek mitigation from a planetary condition such as that of climate change? To what extent is top-down system hierarchy initiating or being influenced by lower levels? As will be seen “emergence” of a nonlinear kind was practiced in two Swedish regions (Västra Götaland and Skåne) under inspection and, more interestingly, learned by path interdependence in a third (Östergötland, centered on Norrköping and Linköping). In brief, one of the two regional agents, Västra Götaland, had by 2001 the outline of a regional eco-innovation strategy, preceding any EU member state, including its own, as well as the EU itself in this so much so that in 2001, it came to be known as the “Gothenburg Model of the Lisbon Strategy” (the EU’s competitiveness strategy). Over approximately a decade, a double feedback loop brought the EU’s advocacy of climate change strategizing back down to regional level in the *Europe 2020* (EU 2010) strategy document. However, long before then in the originating region, regional cluster initiatives inflected toward sustainability had “emerged” as practical actions. Moreover, such regional initiatives were “emergent” elsewhere in the same member state, and the member state itself was becoming more active. Thus such “edge of chaos” regional system adaptability was moving beyond the molecular level due to the exercise of *transversality* as regions and firms sought innovation by stimulating information flow and knowledge appreciation among unlike kinds of cluster. Nevertheless, eventually the EU resource-incentive narrative of “Grand Challenges” emanating from the highest system level gave a further degree of coherence to national and regional strategy discourse, expressing a third feedback loop (or “phase change”) in strategy emergence. A fourth will probably be added when *regional* policy emergence influences the formation of *national* strategy with its own resource-incentive discourse.

Skåne region is committed to giving greater identity and focus to its established and nascent industries by promoting its cluster policy which targets about eight fields. However, regime management builds upon transversal thinking and practice. These recognize the evident advantages of filling regional “white spaces” by stimulating the discovery of “revealed relatedness” and promoting transversal or interface projects and initiatives among clusters. As it stands, the clusters are mostly new and rather weak, except for life sciences, food, and film, but Skåne’s position on the Swedish periphery yet a Scandinavian core, due to its proximity to Copenhagen, means geographic proximity is important, something recognized in the status of the international Medicon Valley life science cluster between Skåne and the Danish capital. In this way, this region operates an “adjacent possible” innovation model inspired by two similar “Grand Challenges” as Västra Götaland in sustainable cities and healthcare but inflected according to regional expertise. Thus, recycling and eco-design are more pronounced elements meeting the national and EU aspirations for a concerted approach to tackling big issues.

Briefly instructive too is the way in which Östsam, Östergötland's regional development agency, and particularly its optoelectronics research institute Acreo, branched away from a 30-year struggle to fit innovative printed electronics technology to a regional and national path dependence upon the packaging products of the pulp and paper industry. A low-intervention, "market-shaping" model here informed the strategy of stimulating the "emergence" of an indigenous supply chain to market the innovation. This failed because it was an overspecialized solution in search of a problem (consignment tracking in the logistics industry) that was already solved by more traditional and cheaper barcode methods. This led to thoroughgoing reversal (phase change) of policy methodology represented in a search for already "emerged" regional industry and clusters customers. These included renewable energy, biotechnology, and healthcare, where potentially appropriate applications of liquid polymer technology might evolve. As Juarrero (2000) observes,

...The precise path that the phase change takes can be explained only after the fact. Such explanation must take the form of a genealogical narrative that reconstructs the bifurcation... Phase changes embody essentially incompressible information... That is why fiction and drama... [are] better than deductions or formulas for explaining... transformations of this sort. (Juarrero 2000, 55)

Apart from our preference for factual over fictional narrative, as a justification for the kind of innovative change management approach explored in this chapter, this is difficult to improve upon. Accordingly, this review presents a rethought and empirically supported base for paying greater attention to the horizontal capacity and bottom-up capabilities of systems to stimulate innovation as an emergent property of interorganizational interaction. Recall this is a rebalancing act that underlines two-way and vertical as well as horizontal feedbacks or phase changes in multilevel process and policy systems.

24.11 Conclusions

Hypothetically, printed electronics began to be rethought once it was realized that its most successful innovative application had emerged in the touch screen controls of smartphones pioneered by *Samsung* of South Korea and early adopters like Taiwan's *HTC*. This looks to be a clear instance of multi-sectoral innovation "blindsiding," arising from technological path dependence since former Nordic leader companies in mobile telephony like *Sony Ericsson* and *Nokia* were locked-in to inferior proprietary and customized telephony system "frames." This can almost perfectly be framed by Mitleton-Kelly's (2011) comparative conclusions on the fate of two hospitals she researched, one that adopted a complexity learning format and one that adopted another approach:

...There was, however, no active learning from these [business process engineering] successes and the focus was very much on attaining financial balance. There was also little active feedback, and few opportunities for staff to get together to review performance and reflect in an open, relaxed and informal atmosphere. Reviewing was done formally in terms

of performance management. By restraining self-organisation and exploration and by not actively reflecting on the outcomes the learning environment was constrained. . . . (Mitleton-Kelly 2011, p. 49)

Hence, we see a revitalized role for learning in RIS Version 3.0. It is that it should be the means whereby innovative organizational change can be motivated against a rather simple, linear model of change based, essentially on cost accounting with little employee engagement, feedback, or learning. Intellectually speaking, this is explicable in complexity science as the failing, cost accounting hospital having, as a system reached the “edge of chaos”:

. . . In complexity theory terms, changes in the ecosystem had pushed the hospital far-from-equilibrium in the sense that they could no longer operate under their existing regime using established norms and procedures. They reached a critical point and had to either do things differently or go downhill. . . . (Mitleton-Kelly 2011, p. 51)

Of considerable influence is that we were able to show from selected examples recently studied that this way of thinking has, partly by a “design” approach broached in more detail in Cooke (2012) how regional innovation systems can be assisted toward optimal outcomes and evolutionary trajectories by utilizing insights from the theoretical material under discussion in this chapter.

References

- Asheim B (1996) Industrial districts as learning regions: a condition for prosperity. *Eur Plan Stud* 4:379–400
- Asheim B, Gertler M (2005) The geography of innovation: regional innovation systems. In: Fagerberg J, Mowery D, Nelson R (eds) *The Oxford handbook of innovation*. Oxford University Press, Oxford, pp 291–317
- Boschma R, Frenken K (2003) Evolutionary economics and industry location. *Rev Reg Res* 23:183–200
- Braczyk H, Cooke P, Heidenreich M (eds) (1998) *Regional innovation systems*. UCL Press, London
- Cooke P (1992) Regional innovation systems: competitive regulation in the new Europe. *GeoForum* 23:365–382
- Cooke P (2007) *Growth cultures*. Routledge, London
- Cooke P (2010) Regional innovation systems: development opportunities from the ‘green turn’. *Technol Anal Strat Manage* 22:831–844
- Cooke P (2012) *Complex adaptive innovation systems*. Routledge, London
- Cooke P, Eriksson A (2011) White spaces innovation in Sweden. VINNOVA, Stockholm
- Cooke P, Heidenreich M, Braczyk H (eds) (2004) *Regional innovation systems*, 2nd edn. Routledge, London
- EU (2010) *Europe 2020*. European Commission, Brussels
- Florida R (1995) Toward the learning region. *Futures* 27:527–536
- Fornahl D, Henn S, Menzel M (eds) (2010) *Emerging clusters*. Edward Elgar, Cheltenham
- Frenken K, van Oort F, Verburg T (2007) Related variety, unrelated variety and regional economic growth. *Reg Stud* 41:685–697
- Geels F (2006) Co-evolutionary and multi-level dynamics in transitions: the transformation of aviation systems and the shift from propeller to turbojet (1930–1970). *Technovation* 26:999–1016

- Harmaakorpi V, Tura T, Melkas H (2011) Regional innovation platforms. In: Cooke P, Asheim B, Boschma R, Martin R, Schwartz D, Tödtling F (eds) *The handbook of regional innovation & growth*. Edward Elgar, Cheltenham, pp 400–410
- Juarrero A (2000) Dynamics in action: intentional behaviour as a complex system. *Emergence* 20:24–57
- Kauffman S (1993) *The origins of order*. Oxford University Press, Oxford
- Kauffman S (1995) *At home in the universe*. Oxford University Press, Oxford
- Kauffman S (2000) *Investigations*. Oxford University Press, Oxford
- Kauffman S (2008) *Reinventing the sacred*. Basic Books, New York
- Martin R (2001) Geography and public policy: the case of the missing agenda. *Prog Hum Geogr* 25:189–210
- Martin R, Sunley P (2010a) The place of path dependence in an evolutionary perspective on the economic landscape. In: Boschma R, Martin R (eds) *Handbook of evolutionary economic geography*. Edward Elgar, Cheltenham, pp 62–92
- Martin R, Sunley P (2010b) Complexity thinking and evolutionary economic geography. In: Boschma R, Martin R (eds) *Handbook of evolutionary economic geography*. Edward Elgar, Cheltenham
- Martin R, Sunley P (2012) Forms of emergence and the evolution of economic landscapes. In: Cooke P (ed) *Reframing regional development*. Routledge, London
- Mitleton-Kelly E (2006) A compl approach to co-creating an innovative environment. *J Gen Evol* 62:223–239
- Mitleton-Kelly E (2011) A complexity theory approach to sustainability: a longitudinal study in two London NHS hospitals. *Learn Org* 18:45–53
- Morgan K (1997) The learning region: institutions, innovation and regional renewal. *Reg Stud* 31:491–503
- Murmann P (2003) *Knowledge and competitive advantage: the coevolution of firms, technology and national institutions*. Cambridge University Press, Cambridge
- Nauwelaers C (2011) Intermediaries in regional innovation systems: role and challenges for policy. In: Cooke P, Asheim B, Boschma R, Martin R, Schwartz D, Tödtling F (eds) *The handbook of regional innovation and growth*. Edward Elgar, Cheltenham
- Rutten R, Boekema F (2007) The learning region: a conceptual anatomy. In: Rutten R, Boekema F (eds) *The learning region: foundations, state of the art, future*. Edward Elgar, Cheltenham, pp 127–142
- Sabel C (2004) Pragmatic collaboration in practice. *Ind Innov* 11:81–87
- Schreyogg G, Höpfel H (2004) Theatre and organisation: editorial introduction. *Org Stud* 25:691–704
- Storper M (2009) Roepke lecture in economic geography – regional context and global trade. *Econ Geogr* 85:1–21
- Sunley P (2011) Worlds of production: conventions and the microfoundations of regional economies. In: Cooke P, Asheim B, Boschma R, Martin R, Schwartz D, Tödtling F (eds) *The handbook of regional innovation and growth*. Edward Elgar, Cheltenham, pp 339–349
- Tödtling F, Trippl M (2011a) Regional innovation systems. In: Cooke P, Asheim B, Boschma R, Martin R, Schwartz D, Tödtling F (eds) *The handbook of regional innovation and growth*. Edward Elgar, Cheltenham, pp 467–481
- Tödtling F, Trippl M (2011b) Regional innovation systems. In: Cooke P, Asheim B, Boschma R, Martin R, Schwartz D, Tödtling F (eds) *The handbook of regional innovation and growth*. Edward Elgar, Cheltenham, pp 455–466
- Tödtling F, Trippl M (2012) Transformation of regional innovation systems; from old legacies to new development paths. In: Cooke P (ed) *Reframing regional development*. Routledge, London, pp 181–192
- Unruh G (2000) Understanding carbon lock-in. *Energy Policy* 28:817–830
- Verganti R (2006) Innovating through design. *Harvard Business Review*, Dec 2006, Reprint R0612G
- Weick K (1995) *Sensemaking in organisations*. Sage, London

Frank G. van Oort and Jan G. Lambooy

Contents

25.1	Introduction	476
25.2	Knowledge Creation and Diffusion	477
25.3	Mechanisms of Knowledge Production and Diffusion in Cities	480
25.4	Agglomeration, Variety, and Pecuniary External Effects	483
25.5	Knowledge Spillovers in the Urban Agglomeration Literature	484
25.6	Conclusions	486
	References	487

Abstract

This chapter provides an overview of current theories and empirical research on cities and the knowledge economy. Two recent and interrelated streams of literature are discussed: the first focusing on agglomeration economies related to increasing returns and knowledge spillovers of firms in cities and the second highlighting the role of knowledge workers and creativity in identifying new and innovative growth opportunities in cities. We argue that analyses using knowledge production functions to capture knowledge flows in cities do not, as of yet, provide true insight into the generation and transfer of different kinds of knowledge. Only recently are various conceptualizations of distance and knowledge transmission channels able to address the heterogeneity of the actors and processes involved in capturing the respective role of cities in knowledge creation. We conclude that the mechanisms that create and diffuse knowledge in cities should be better embedded into both streams of literature. The current discourse on agglomeration externalities obviously needs such conceptual and

F.G. van Oort (✉) • J.G. Lambooy

Department of Economic Geography, Faculty of Geosciences, Utrecht University, Utrecht,
The Netherlands

e-mail: f.g.vanoort@uu.nl; J.G.Lambooy@kpnmail.nl

methodological views to address current impasses. In particular, evolutionary economic geographical concepts are promising in explaining the innovative behavior of growing firms and organizations in cities, carefully addressing the heterogeneity of the actors involved, spatial scale, selection and survival, as well as time and path dependency.

25.1 Introduction

In the 1980s, many people were convinced that cities as centers for social and economic dynamism were disappearing. Wealthier people wanted to live outside the cities, in larger villages, or in suburban areas. The widespread use of the car and the rise of information and communication technologies (ICT's), as well as the concentration of the socially and economically disadvantaged, made many observers think that we would see a complete transformation of communication, spatial configurations, and social and economic structures. The end of distance and spatial concentration seemed near (Gaspar and Glaeser 1998). The current view is completely different; today, views on urban development hold that distance still matters and that urban concentration still continues (Glaeser 2011). At first glance, it seems easy to understand why cities are increasingly the preferred mode of human settlement. They save on infrastructure, reduce trade costs, and enhance interaction. In developing countries, there's nothing particularly new about urbanization as an expression of development. At second glance, however, cities are associated with many costs – land, pollution, and other externalities. We once believed that suburbanization combined the primary advantages of urbanization with these lower land costs and other externalities, but density has returned to many parts of the world.

Examining the arguments behind this modern urbanization view, we find two major approaches. The first is based on the theory of agglomeration economies with increasing returns and easy access to knowledge (Jacobs 1984; Krugman 1995), and the second is based on the idea that (larger) cities are strong because they claim to be the physical concentration of skilled knowledge workers and the creative class (Glaeser 1999; Florida 2002). Both approaches lead to the hypothesis of an expected higher labor productivity. This raise of productivity seems to be the case, although the explanation can differ. In Europe, the largest urban areas, in particular the London-Randstad-Paris-Frankfurt-Milan axis, contribute much more to their national GDPs than could be expected judging by their population sizes (Ciccone 2002). The same conclusion was reached by Glaeser (2011) for the USA.

The question that can be raised is whether knowledge is a decisive factor in explaining (a) higher productivity and (b) the stronger attraction of knowledge workers, or more general, of the creative class. In urban economics, knowledge receives increasing attention as a source of growth. Apart from knowledge, urban growth is explained by many other variables, and by the concentration of creative people, with the expected concomitant development of new knowledge and innovations (Combes et al. 2008). Urban economics focus on properties associated with

agglomeration advantages, external economies of scale, increasing returns, and the development of a variety of producers and consumers. Knowledge, its generation, and its use in innovations are perceived as the principal variables. However, the concept of knowledge in itself is not entirely clear, and neither are the mechanisms of its impact on productivity. It can be seen as a separate production factor or as an attribute, in one way or another, linked to capital goods and labor. Most economic studies emphasize the second interpretation. In endogenous economic growth theory, knowledge is seen as an output of investment in research and development (R&D). This kind of investment can be defined more broadly, as knowledge-intensive inputs, such as in new capital goods and new labor, to increase R&D. To make things more difficult, in the aforementioned case, knowledge can be both an input and an output.

In this chapter, a (non-exhaustive) survey of theories and empirical applications on research focused on cities, knowledge, and innovation is provided. We structure the theories using two recent and interrelated streams of literature: the first focusing on agglomeration economies related to increasing returns and knowledge spillovers of firms in cities and the second highlighting the role of knowledge workers and creativity in identifying new and innovative growth opportunities in cities. [Section 25.2](#) provides an overview of historical and current conceptualizations of knowledge and knowledge diffusion. [Section 25.3](#) provides a discussion on knowledge production function methodologies applied to cities, and concludes, consistent with [Sect. 25.2](#), that future research should increasingly and explicitly focus on the transfer mechanisms of knowledge diffusion. [Section 25.4](#) focuses on the related literature of agglomeration externalities and its link to innovation and urban economic growth. [Section 25.5](#) confronts the central plea for a better embedding of knowledge transfer mechanisms in agglomerations with current agglomeration discourse and methodologies. Conclusions on new and necessary conceptual and methodological views are presented in [Sect. 25.6](#).

25.2 Knowledge Creation and Diffusion

A useful distinction is the one in “data” (facts or unstructured information), “information” (structured data and standardized knowledge) that can be easily transferred via markets, “practical knowledge” (dispersed over economic actors and belongs to particular individuals) that can be used in commercial activities or applied directly in production processes, and “scientific knowledge” that can be applied after further research and development. Knowledge is acquired through the interactive process of learning, based on the cognitive competencies and experiences of the actors. A distinction can be made between “tacit” and “codified” knowledge. Tacit knowledge can be defined as the person-specific knowledge that people have developed through a process of learning-by-doing or learning-by-using, meaning that a person is able to use it but is not, or is not yet, able to codify it in a transferable form, such as a book, patent, or a mathematical formula.

It has to be emphasized that these kinds of knowledge are not fixed. Nonaka and Takeuchi (1995) have emphasized that tacit knowledge and codified knowledge have to be perceived as dynamic attributes that can be transformed into one other. After a period of application, tacit knowledge can be codified and, vice versa; sometimes, codified knowledge can be developed further and can become tied to individuals. The ways in which learning and the transfer of knowledge across organizations and in spatial settings can be investigated and measured have triggered much debate and research.

Codified knowledge (like patents and books) can be transferred to other users and is most often sold in markets. Tacit knowledge is person and context specific; hence, it has to be transferred or sold connected to a person, as with labor mobility. In some instances, knowledge is transferred for free, as in certain teaching situations or in networks, based on friendship. Tacit knowledge can develop into codified knowledge. This happens when, through research, certain skills, feelings, or capabilities are formalized or defined by rules, as has been performed with chess programs and in medical research. The reverse also occurs when codified knowledge becomes “embedded” in behavior and becomes tacit through its implementation in new situations, through experience, or by sharing within a group. Recent questions on the subject involve the possibilities of transferring tacit knowledge not only via interpersonal contacts but also via modern tools such as TV, the Internet, and mobile phones. Another question is whether the use of knowledge for economic goals and its transfer are related exclusively to production and consumption. Additionally, the question arises regarding whether wider opportunities to increase welfare emerge, for instance, in the arts, or whether measures to increase the sustainability of cities are implemented (Glaeser 2011). Finally, the link between knowledge, skills, and creativity is important. The use of knowledge and, in a broader sense, creativity is not always directly related to economic goals of production and consumption. Creativity can be used for solving personal or family problems, or expressed in the arts. However, this also can lead to higher satisfaction or even to the rise of certain forms of consumption, through which the production of goods and services can increase. There are also many indirect links between creativity, knowledge development, and economic growth.

In endogenous growth theory, the generation of new knowledge and innovations is explained by increased investment in R&D. This concept was also described in Schumpeter's 1942 book, in which he saw the entrepreneurial function of management replacing the risk-taking and innovative individual hero-entrepreneur. In this view, it is possible to create a knowledge production function (KPF), with investments in R&D as input and knowledge and patents as an output. Recent applications at the regional and urban level show that this line of thinking is fruitful for explaining urban growth differentials (Acs 2002; Breschi and Lissoni 2009). However, in this KPF approach, the important other causality, the line starting from the demand side, is often neglected. The failure to meet the preferences and needs of users has been consistently stressed as a major cause of unsuccessful R&D for over 30 years. This is especially important in urban agglomerations, as is emphasized in studies on consumer cities. In modern urban agglomerations,

consumers are increasingly seen as a source of changing demand and new ideas (Glaeser et al. 2001). However, it is not yet clear how the lines of knowledge generation and application can be connected to producers and consumers in urban contexts. It is particularly necessary to investigate the relative importance of markets (prices) and formal and informal networks as carriers of knowledge. Krugman (2009, p. 567) also emphasized the unclear role of spillovers: “it’s not at all clear how to think about the spatial limits of spillover. Do you have to be in the same city to reap positive externalities from other producers in the same industry?” However, for cities with a wide variety of different kinds of economic activities and international relations, this question becomes even more important. The analysis incorporating production functions does not provide us with insight of the generation and transfer of different kinds of knowledge (yet). The investigation of knowledge in cities, as well as other kinds of knowledge, such as scientific knowledge, practical knowledge, and Polanyi’s tacit knowledge, is in need of receiving greater attention. The generation and diffusion of these heterogeneous kinds of knowledge can vary and need to be considered as having different effects on urban economic variables. In turn, urban contexts can be perceived as having effects on the generation and diffusion of the various kinds of knowledge. Cities can cumulatively specialize in certain kinds of knowledge connected to different economic specializations (Duranton and Puga 2005).

It is generally accepted that geographical contexts, such as agglomerations, influence the generation and dissemination of knowledge, although other mechanisms besides markets are not always included in explanations of urban development. In urban economics, the role of geographical distance as such is reflected in transport costs (both in the old and in the new economic geography), in the impact on social relations, and in the availability of knowledge embedded in labor, as in the popular view on the development of industrial clusters. In theories of urban hierarchy, distance costs are seen as a decisive factor explaining the differences of the quality of services and amenities in centers of different size. Distance is generally translated in terms of costs or in missed opportunities, stemming from the failure to note chances to produce, sell, or purchase goods or services.

Boschma (2005) emphasized that the word “distance” can be interpreted in different ways. Geographical distance is not the only important factor; social and cultural distance can also be important. This distinction is especially relevant in analyzing the geographical sources of knowledge and innovation because in a globalizing world many researchers have connections with people in other places. Nevertheless, it seems that distance costs and perception barriers are important factors in the analysis of agglomeration advantages because closer distance, no matter how it is measured, seems to foster the development of knowledge and innovations (Breschi and Lissoni 2009). One of the strong attributes of closer distances is the easier transfer of knowledge. This is related to the uncertainties that are inherent to the economic process and to the rapid changes that knowledge development has shown in our time. Such changes in knowledge cause the need for continuous adaptations by economic actors and hence the tendencies to locate

generators of knowledge locally, such as in universities and other R&D organizations (Audretsch and Feldman 1996).

The dynamic properties of knowledge are associated with various forms of dissemination and with their applications as innovations, with concomitant changes in supply and demand conditions. Through this process, continuously disturbing effects on equilibrium exist. Markets have to respond quickly to rapidly changing contexts. Metcalfe (1998) even argued that capitalism is restless because knowledge is restless. Innovations can also change market structures by creating new monopolies, as in the case of radical new technologies, especially GPTs (general purpose technologies).

The consequences of the application of knowledge for innovations are important. However, it is also important to examine the sources of knowledge (or its generation), where the cognitive attributes of economic actors and the organizational capacities of entrepreneurs are decisive properties of a knowledge-enhancing urban society. Economic actors vary considerably in their cognitive capacities and other attributes. This heterogeneity is one of the strengths of cities and an important reason for their continuous creativity. Situations involving heterogeneous actors and conditions lead to continuous uncertainties, which offer “gaps in information” for the existent markets, and this, in turn, can give entrepreneurs opportunities for innovative actions. This is particularly the case when new knowledge leads to pecuniary external effects (Metcalfe 1998).

25.3 Mechanisms of Knowledge Production and Diffusion in Cities

Knowledge and innovation are closely associated. Knowledge is not just “given” but has to be generated. Schumpeter (1934) emphasized that the generation of knowledge is the result of a process of wider social significance. The generation of knowledge is important for the development of the quality of labor and capital goods and is decisive for innovation and economic growth. However, the direction of the causality of the relationship between knowledge, innovation, and economic growth is not always clear. Knowledge can be the result of investment in growing sectors of the economy, but it can also be developed by people with inquisitive minds, without the purpose of commercial application. Knowledge and innovation do not only start at the supply side with research and investment. Rather, the causality may sometimes start with changing demand because certain cultural developments alter the structure of demand through concomitant changes in sectoral structures. This can also happen if wealth increases due to the expansion of foreign markets with already-existing kinds of products. All types of change can lead to the demand of new knowledge, products, and production technologies. This section provides a discussion of production function methodologies and concludes by recognizing the need to explicitly focus future research on the transfer mechanisms of knowledge diffusion. The argumentation is neo-Schumpeterian in character.

Being that knowledge is hard to appropriate, as we argued in the previous section, it generates benefits to other agents through several spillover mechanisms. Understanding the geographical structures that underlie these spillover benefits is necessary for any evidence-based innovation policy to stimulate a region's (or collection of regions, such as Europe) transformation toward a knowledge-based society. Recent years have seen many macro studies on the effect of knowledge spillovers on innovation. Such studies generally apply a knowledge production function to differentiate regional innovation outputs from regional knowledge inputs, as well as from knowledge spillovers from other regions. The strength of interregional knowledge flows is generally assumed to decrease rapidly with geographical distance (Acs 2002), while others have attempted to measure spillovers directly by patent citations (Breschi and Lissoni 2009). Despite the fact that previous research has produced a certain degree of empirical coherence (Fritsch and Slavtchev 2007), it has proven difficult to distinguish between different channels of knowledge spillovers; this has subsequently led scholars to rely on specifications that are suggestive of knowledge spillovers without explicitly modeling the mechanisms through which they occur in practice. Understanding the mechanisms that are behind knowledge spillovers is obviously of the greatest importance for designing effective innovation policies.

The step forward to be taken in empirical research on knowledge production and diffusion in cities and city regions is to correctly model spillover mechanisms of knowledge (correctly). This means that, conceptually speaking, one should take into account via a single framework both geographically localized knowledge spillovers (by examining the extent to which regions profit from regions knowledge inputs nearby) and knowledge spillovers stemming from research collaborations (by examining the extent to which regions profit from network connections to other regions). Put differently, one can analytically distinguish between the "space of place" creating geographically localized knowledge spillovers and the "space of flows" creating spillovers in global networks (Castells 1996). In contrast to geographically localized channels of knowledge spillovers, such as spin-off dynamics and informal networking (Breschi and Lissoni 2009), two promising networked channels of knowledge diffusion come to the fore: research collaborations between firms and universities that are increasingly taking place over longer distances and the mobility of globally operating knowledge workers.

The presence of both private and public research organizations, such as universities and the laboratories of multinational corporations, is generally assumed to have a large impact on urban innovation due to their ability to attract knowledge workers and generate localized knowledge spillovers resulting from their research (Acs 2002). Various empirical studies have suggested the presence of localized academic knowledge spillovers for the USA and various European countries. It has been stressed that highly skilled workers can be regarded as carriers of knowledge diffusion and key drivers of regional innovation and growth. Individuals impact knowledge diffusion through two main and complementary channels: on the one hand, their ability to move from one place or one organization to another and, on the other hand, their ability to enter networks. The mobility patterns seem to be

predominantly local, though some evidence indicates that, due to the intrinsic universality of science, researcher's labor markets – in particular academic researchers' mobility – tend to be more international than other labor markets. Motivations for labor mobility could be related to scientific, economic, cultural, and personal factors. It is expected that nonpecuniary scientific factors, such as the quality of the university, the availability of research budgets, personnel and material research infrastructure, and institutional reputation, are more important than economic factors (Jons 2007). Due to the rising demand of specialized scientific labor, a reputation for scientific quality and openness is a critical factor for attracting excellent researchers. However, differences in wage levels, career opportunities, and concentration effects (talent is attracted by talent) are also thought to be relevant. There also exists a relation to soft factors, for example, language, cultural affinity, living environment, and personal motives (Florida 2002). In line with these insights, many countries have implemented regional innovation policies based on the presence of universities and research institutes in a city or region. In particular, initiatives have been set up to attract skilled workers and facilitate their movement. Systematic research on this kind of labor mobility and migration is lacking.

Besides the importance of local labor markets and spin-off dynamics, a growing body of research stresses the role of networks between individuals and between organizations as mechanisms for knowledge spillovers. Informal networking often takes place at the regional level and, as a result, knowledge spillovers are localized to the extent of these networks. Formal networks of research collaboration are an additionally important mechanism of knowledge spillovers; however, empirical research on the spatial dimension of these networks has suggested that they largely occur at the national or even international scale (Ponds et al. 2010; LeSage et al. 2007). The structure of collaboration networks thus needs to be taken into account to fully understand the impact of researchers' knowledge spillovers (Barber et al. 2011). Despite the increase in literature analyzing knowledge flows in different organizations and geographical contexts, little is known about actual knowledge circulation and its impact on community, as well as urban and regional knowledge creation, diffusion, and quality. At least two weaknesses in the existing studies stand out. First, most studies on localized knowledge spillovers claim that knowledge does not circulate freely across regional boundaries because it is tacit; on the other hand, these studies remain elusive on the specific mechanisms of diffusion of tacit knowledge (Audretsch and Feldman 1996). Studies of knowledge diffusion, with some exceptions, tend to focus only on codified forms of knowledge and on formal channels of transmission (i.e., patents, patent citations, publications, R&D). Second, studies on migration and mobility offer important insights into the motivational factors behind the decision of scientists to move. However, this literature is mostly based on anecdotic or only qualitative evidence. Quantitative evidence instead focuses only on selected groups of skilled workers (e.g., graduates, star scientists) and is limited to cases by country. Future research should shift attention from codified toward more tacit forms of knowledge and from stocks of knowledge toward flows and networks of knowledge.

25.4 Agglomeration, Variety, and Pecuniary External Effects

Agglomeration advantages have been connected with increasing returns to scale, external effects, and with the variety of producers and consumers (Combes et al. 2008). New knowledge changes markets, market structures, and production technologies (including organizational structures). Cities can be perceived not only as locations with agglomeration advantages but also as locations of interacting producers and consumers. Interaction occurs not only via markets but also via social networks, sometimes indicated with “buzz” (Storper and Venables 2004). Baumol (2002) argued that markets are predominant in the process of knowledge creation and diffusion, even when other mechanisms of dissemination also exist, such as social networks, labor mobility external effects (externalities), spillovers, and spin-offs. In various publications, the transfer of knowledge is assumed to be different from other goods and services, most often because it is perceived as a public good, freely accessible to anyone, or as given. In traditional growth accounting literature, “technological knowledge” was accepted as exogenous. Most often, the transfer of knowledge, especially codified knowledge such as books and patents, has a price and occurs via markets. This is also the case for labor mobility, where a higher wage is paid to newly attracted experts. The prices do not always completely cover the costs created by the generators of knowledge; subsequently, in many cases, we can highlight market failure or even more pecuniary external effects. In other cases, one could even address unpaid positive external effects; Alfred Marshall indicated this latter effect by asserting that “it is in the air.”

In the case of technological development and expanding or changing demand, new opportunities are created. This can change the relations between firms in different sectors, which are all confronted with an expanding or changing supply and demand. This requires new production methods and products, a good base for an endogenous process of increasing demand for new knowledge and innovations. Adam Smith and Alfred Marshall emphasized the interrelations between firms with the division of labor and industrial districts. In cities, changes could have their origin in new (technological) knowledge, increasing wealth, or new trade relations, leading to new supply and demand and hence to a larger economic base. As Adam Smith said, “the size of a town depends on the size of the market.” The interrelations between trade, the differentiation of consumers, different kinds of firms, and the development of increasing returns have experienced a resurgence in new trade economics and new economic geography. The economic process is influenced by these factors, but in cities, we can also observe the special influence of positive and negative external effects, which in economic theory has been defined as market failures. However, in the case of increasing market size (growing demand within cities and regions through increasing wealth and trade effects) positive pecuniary external effects are related to increasing returns. This leads to the conclusion that pecuniary external effects are not merely market failures but dynamic opportunities for innovators. Allyn Young (1928), then president of the American Economic Society, emphasized the positive impact of the interrelatedness of firms in the production process. In this structure of relations, as well as in the case of the

development of new technologies, pecuniary external effects could develop because entrepreneurs detect new opportunities for higher profits with new products or new technologies. He stressed that in this case, the opportunities to invest in new technology could result in pecuniary external effects and in increasing returns to scale for the entire structure of related firms. Pecuniary external effects offer new opportunities for entrepreneurs by creating “gaps” in the market. Entrepreneurs can establish new firms and produce new products (physical goods or services). Schumpeter (1934) emphasized that innovation is related to this entrepreneurial function of seeing the gaps in the relation of demand and supply by establishing new firms and improving the allocation of resources.

25.5 Knowledge Spillovers in the Urban Agglomeration Literature

Despite this complex and nuanced way of conceptually linking innovation – the introduction and the application of new or existing knowledge – with growth and cities, an ever-growing body of empirical literature on urban externalities remains rather inconclusive on the exact agglomeration circumstances that optimally enhance growth and innovation in cities. In such literature it is argued that externalities or spillovers occur if an innovation or growth improvement implemented by a certain enterprise increases the performance of other enterprises without the latter benefiting enterprise having to pay (full) compensation. Spatially bound externalities are related to enterprise’s geographical or network contexts and are not related to internal firm performance. All discussions of spatial externalities can be linked to a twofold classification in which the sources of agglomeration advantages are grouped. *Localization economies* usually take the form of Marshallian (technical) externalities whereby the productivity of labor in a given sector of a given city is assumed to increase with total employment in that sector. In short, they arise from labor market pooling, the creation of specialized suppliers, and the emergence of technological knowledge spillovers. The strength of local externalities is assumed to vary, such that they are stronger in some sectors and weaker in others. The associated economies of scale comprise factors that reduce the average cost of producing commodities. External scale economies are applicable when the industry to which the firm belongs (rather than the firm itself) is large. An urban system is composed of (fully) specialized cities, provided that the initial number of cities is large enough; such systems occur contingent on further assumptions on crowding (congestion costs that increase with population triggers dispersion), perfect product, and labor mobility within and between locations, not to mention the influence of large agents. Once cities exist, *urbanization economies* that apply to all sectors become equally important. Urbanization economies are often interchangeably mentioned with Jane Jacobs’ diversity externalities, as (sectoral) diversity tends to be larger in cities than outside them. Frenken et al. (2007) showed that a distinction between variety and diversity externalities and urbanization economies is necessary. A large body of empirical literature has grown around testing

these types of externalities in relation to knowledge spillovers using sectoral specialization, sectoral diversity, and density data from cities. The assumption is that if knowledge spillovers are important to growth and firm dynamics, they should be more easily identifiable in cities where many people are concentrated into a relatively small and confined space where knowledge is transmitted more easily. This literature has evolved in a rather polarized discussion on the question of whether sectoral specialization (clusters) or sectoral diversity matter for economic growth and innovation in cities. Three recent meta-analyses and overviews clearly show the limitations of this empirical approach (De Groot et al. 2009; Melo et al. 2010; Beaudry and Schiffauerova 2009); the outcomes of the many empirical analyses using the Glaeser et al. (1992) framework on agglomeration externalities appear to be highly dependent on spatial scale, sectoral detail, time frame, institutional context, and the construction of indicators and variables. Twenty years of research have not convincingly answered the question “Who was right, Marshall or Jacobs?” (Beaudry and Schiffauerova 2009). The answer is ambiguous; both specialization and diversity are related to growth in different aspects and ways.

In principle, this answer is rather unsatisfactory scientifically for understanding the relation between urban growth, spillovers, and innovation. It is very plausible that the prevailing static urban economic modeling approach, confronting the Marshallian versus the urbanization externalities approach, falls short both conceptually and methodologically, and in its present form is unable to test this important issue satisfactorily. In its conceptual sense, this was previously noted by Lambooy and Van Oort (2005), who suggest four heterodox aspects attached to urban and regional economic growth that are currently (still) relatively unaccounted for in research and should be taken more seriously. These are (a) the importance of the life stages and time frames of firms, technologies, and sectors, or development paths of firms, sectors, cities, and systems of cities; (b) specific spatial networks not showing (clear) relations to the forces of contiguous economic agglomeration; (c) specific urban and regional factors explaining why and through which transmission channels agglomeration forces influence sectors and firms differently, depending on the period of economic development and the various technological trajectories; and (d) factors related to forces that cannot be explained using equilibrium approaches, for example, the relation with institutional structures, path-dependent development, the way selection works out for new technologies and firms, innovation, the rise of new technologies and new regional concentrations of firms, spillover mechanisms, and (co-)evolution. These four heterodox aspects of economic theory and empirics are attached to evolutionary economic development trajectories in a wide range of cities, regions, and countries in the same manner (McCann and Van Oort 2009). In the evolutionary geographical research tradition, much more emphasis is placed on the interaction of the relevant urban and regional environment, with locational choices being made by individual firms and investors (Boschma and Martin 2010). In these traditions, a strong preference exists to allow for the differentiation of firms and types of behavior and locations, addressing the heterogeneity in actors and innovation in cities that were signaled in the previous section. The concept of related variety, indicating that successful sectors in regions diversify over time, though mostly in relation to existing competences and

specializations, is an important exponent of new conceptualizations in the agglomeration, innovation, and growth discussion (Frenken et al. 2007).

Together with these conceptual issues, methodological issues arise as well. More emphasis on a firm or consumer's personal agglomeration circumstances requires a modeling approach that takes firms and consumers as starting point. Duranton and Overman (2005) and Combes et al. (2008) argue that many measures of concentration use arbitrary spatial units (such as provinces, municipalities, or postcodes), which may be problematic, as they may lead to biases. Continuous space specifications of agglomeration circumstances of individual firms therefore become more important in present and future research, avoiding the problems of modifiable areal units signaled in the three review articles. Furthermore, issues of causality, endogeneity, selection, and sorting have to be addressed more thoroughly to ensure that the econometric analyses produce reliable outcomes. Both conceptual and methodological renewal are needed to investigate the nature and origin of knowledge creation and diffusion (transfer mechanisms and absorptive capacity of actors), addressing the central issue of heterogeneous actors as well as the varying contexts in the organization of sectors and networks.

25.6 Conclusions

We have provided an overview of historical and current conceptualizations of knowledge, knowledge diffusion, and innovation in cities. We have argued that knowledge is based on processes of learning as well as research and development, that it is both person specific as well as context specific, and that it can be codified and included in the quality of capital and labor. Recently, much empirical research has focused on the creation and transfer of knowledge across organizations in spatial contexts. We argue that analyses using knowledge production functions to capture these flows generally do not provide us with true insight into the generation and transfer of different kinds of knowledge. Only recently have various conceptualizations of distance and knowledge transmission channels been empirically related to knowledge creation and diffusion, addressing heterogeneity in related actors and processes, and capturing the role of cities in them. Our discussion of knowledge production function methodologies applied to cities has concluded that future research should increasingly and explicitly focus on the transfer mechanisms of knowledge diffusion. This is especially true for research on the mobility of (star) knowledge workers and on the evidently fruitful collaborations between firms and universities. To incorporate this in empirical modeling, econometrical knowledge and innovative applications are needed in this field of research. The chapter has further argued that markets remain the most important kind of interaction for economic actors, even in the case of knowledge. This nuances the large focus on nonmarket factors as put forward in the growing literature on urban competitiveness and innovation. We confronted the plea for a better embedding of the mechanisms that create and diffuse knowledge in agglomerations with current agglomeration discourses and methodologies. We conclude that to address the

apparent impasse on the measurement and interpretation of agglomeration externalities, new conceptual and methodological views are needed here as well. In particular, evolutionary economic and geographical concepts are promising for explaining the innovative behavior of growing firms and organizations in cities, carefully addressing the heterogeneity in the actors involved, spatial scale, selection and survival, and time and path dependency. For this, accompanying econometric tools have to be applied, such as continuous space modeling and causality analysis. The future of urban agglomeration research is thus in the interplay of conceptual and methodological renewal, in close relation to already-established insights: what is needed is renewed and related variety in conceptualization and testing.

References

- Acs ZJ (2002) Innovation and the growth of cities. Edward Elgar, Cheltenham
- Audretsch DB, Feldman MP (1996) R&D spillovers and the geography of innovation and production. *Am Econ Rev* 86:630–640
- Barber MJ, Fischer MM, Scherngell T (2011) The community structure of research and development cooperation in Europe: evidence from a social network perspective. *Geogr Anal* 43:415–432
- Baumol WJ (2002) The free-market innovation machine. Princeton University Press, Princeton
- Beaudry C, Schiffauerova A (2009) Who's right, Marshall or Jacobs? The localization versus urbanization debate. *Res Policy* 38:318–337
- Boschma RA (2005) Proximity and innovation. A critical assessment. *Reg Stud* 39:61–74
- Boschma R, Martin R (eds) (2010) Handbook of evolutionary economic geography. Edward Elgar, Cheltenham
- Breschi S, Lissoni F (2009) Mobility of skilled workers and co-invention networks: an anatomy of localized knowledge flows. *J Econ Geogr* 9:439–468
- Castells M (1996) The rise of the network society. Blackwell, Oxford
- Ciccone A (2002) Agglomeration effects in Europe. *Eur Econ Rev* 46:213–227
- Combes PP, Mayer T, Thisse JF (2008) Economic geography. The integration of regions and nations. Princeton University Press, Princeton
- De Groot HLF, Poot J, Smit MJ (2009) Agglomeration externalities, innovation and regional growth: theoretical perspectives and meta-analysis. In: Nijkamp P, Capello R (eds) *Handbook of regional growth and development theories*. Edward Elgar, Cheltenham/Northampton, pp 256–281
- Duranton G, Overman HG (2005) Testing for localization using micro-geographic data. *Rev Econ Stud* 72:1077–1106
- Duranton G, Puga D (2005) From sectoral to functional urban specialisation. *J Urban Econ* 57(2):343–370
- Florida R (2002) The rise of the creative class. Basic Books, New York
- Frenken K, van Oort FG, Verburg T (2007) Related variety, unrelated variety and regional economic growth. *Reg Stud* 41:685–697
- Fritsch M, Slavtchev V (2007) Universities and innovation in space. *Ind Innov* 14:201–218
- Gaspar J, Glaeser EL (1998) Information technology and the future of cities. *J Urban Econ* 43:136–156
- Glaeser EL (1999) Learning in cities. *J Urban Econ* 46(2):254–277
- Glaeser EL (2011) Triumph of the city. Penguin Press, London
- Glaeser EL, Kallal HD, Scheinkman JA, Schleifer A (1992) Growth in cities. *J Polit Econ* 100:1126–1152
- Glaeser EL, Kolko J, Saiz A (2001) Consumer city. *J Econ Geogr* 1:27–50

- Jacobs J (1984) Cities and the wealth of nations. Random House, Toronto
- Jons H (2007) Transnational mobility and the spaces of knowledge production: a comparison of global patterns, motivations and collaborations in different academic fields. *Soc Geogr* 2:97–114
- Krugman P (1995) Development, geography, and economic theory. The MIT Press, Cambridge, MA
- Krugman P (2009) The increasing returns revolution in trade and geography. *Am Econ Rev* 99:561–571
- Lambooy J, van Oort FG (2005) Agglomerations in equilibrium? In: Brakman S, Garretsen H (eds) Location and competition. Routledge, London, pp 79–108
- LeSage J, Fischer MM, Scherngell T (2007) Knowledge spillovers across Europe: evidence from a Poisson spatial interaction model with spatial effects. *Pap Reg Sci* 86:393–421
- McCann P, van Oort FG (2009) Theories of agglomeration and regional economic growth: a historical review. In: Capello R, Nijkamp P (eds) Handbook of regional growth and development theories. Edward Elgar, Cheltenham, pp 19–32
- Melo PC, Graham DJ, Noland RB (2010) A meta-analysis of estimates of urban agglomeration economies. *Reg Sci Urban Econ* 39:332–342
- Metcalf JS (1998) Evolutionary economics and creative destruction. Routledge, London
- Nonaka I, Takeuchi H (1995) The knowledge-creating company. Oxford University Press, New York
- Ponds R, van Oort FG, Frenken K (2010) Innovation, spillovers, and university-industry collaboration: an extended knowledge production function approach. *J Econ Geogr* 10:231–255
- Schumpeter JA (1934) Theory of economic development. MIT-Press, Cambridge, MA
- Storper M, Venables AJ (2004) Buzz: face-to-face contact and the urban economy. *J Econ Geogr* 4:351–370
- Young A (1928) Increasing returns and economic progress. *Econ J* 38:527–542

Emmanouil Tranos

Contents

26.1	Introduction	490
26.2	Networks and Associated Concepts	491
26.3	Knowledge Networks in a Knowledge-Based Economy	494
26.4	Innovation Networks and Different Types of Proximity	497
26.5	Innovation Networks: Some Methodological Approaches	500
26.6	Conclusions	502
	References	503

Abstract

This chapter reviews the importance of networks in the innovation process from a spatial perspective. Such networks are part of different scale systems of innovation and are essential to the creation of knowledge externalities. It is well established in the extant literature that innovation does not occur in isolation, and furthermore, interorganizational networks facilitate innovation creation. Social networks, trust, and local embeddedness play key roles in the formation of such networks. In addition, relational perspectives, such as non-geographical proximities, are also vital factors for the creation of innovation networks, the main objective of which is knowledge creation. Important enough, the latter can be approached as crucial production factor in the frame of the knowledge economy. Moreover, scale is an important attribute of such networks, as both local and global links are important in the innovation process.

E. Tranos

Department of Spatial Economics, VU University, Amsterdam, The Netherlands
e-mail: e.tranos@vu.nl

26.1 Introduction

This chapter aims to review the importance of network formations in the innovation process adopting a spatial perspective. The starting point for this journey is the systemic understanding of the innovation process: innovation does not occur in isolation but is the outcome of systemic interactions among various actors. The spatiality of such systems of innovation (SI) and more specifically networks of innovation is the main focus of this chapter. Unavoidably, such an analysis is underpinned to some extent by evolutionary processes, which capture the change of organizations, institutions, and their ties over time. Importantly, SI highlight the role of interactions between actors in the innovation process and the related feedback mechanisms. Thus, innovation activity is a collective process, which is based on actors interacting together to transfer, exchange, and create knowledge (Edquist 1997). Such knowledge is a necessary input for the *innovation* production. It needs to be stressed here that despite the great interest from academics and policy makers in understanding the mechanisms behind innovation creation, there is still not a commonly accepted definition for innovation. For the needs of this chapter, the following definition is adopted: “An innovation is the implementation of a new or significantly improved product (good or service), or process, a new marketing method, or a new organizational method in business practices, workplace organization or external relations” (OECD and Eurostat 2005, p. 46).

The systemic understanding of the innovation process was initiated by the introduction of the concept of national innovation systems (NIS). Pioneers in this conceptual formation were Christopher Freeman and Bengt-Åke Lundval, who introduced the idea that different innovation actors are linked together forming a knowledge exchange network structure. Such a systemic understanding of innovation is based on three components: (a) the actors of the innovation system, which are private (individuals and firms) and public organizations (universities or state authorities), the affiliation of which with the SI defines the boundaries of the system; (b) the relations and interactions among these actors for knowledge creation and diffusion; and (c) the attributes of the actors.

Soon after the systemic view on innovation was established, this approach was then transferred to a regional scale known as the regional innovation system (RIS) (Cooke 1992). This scalar change came as a response to empirical evidences that innovative activity is neither uniformly nor randomly distributed across geographical space. The main explanation is that knowledge, especially tacit – that is, the non-codified part of knowledge – can only be transferred via face-to-face interactions and within an atmosphere of trust. Thus, geographical proximity can act as a facilitator for the innovation production. However, as it will be highlighted later in this chapter, geography is not the only proximity dimension useful in this context. During 1990s, the RIS concept was developed as an analytical framework for policy makers and academic researchers to understand the innovation process in regional economies (Cooke 1992). RIS focuses on interactive networks of nodes

and linkages on a regional scale to understand a city or region's innovation performance. Thus, the spatiality of interactions among innovation actors becomes an important element of the innovation production.

The last point leads the discussion to the core of this chapter. Innovation is a systemic process, which is facilitated by networks of actors. The latter is a systemic representation of interactions among innovation actors. However, such systems are anything but *a-spatial* as spatial and territorial dynamics facilitate interactions and directly affect innovation actors. After analyzing the systemic nature of innovation in the introduction section, this chapter then focuses on further understanding innovation networks. Then, [Sect. 26.3](#) highlights the importance of networks in knowledge creation and in the frame of the knowledge economy. [Sect. 26.4](#) focuses on the impact of different proximities in network formations, and [Sect. 26.5](#) reviews the main current methodological challenges in modeling spatial innovation networks. Finally, the chapter ends with a conclusion section.

26.2 Networks and Associated Concepts

Owing to Cooke and Morgan's ([1993](#)) seminal work on the network paradigm, the importance of networks in regional economic growth is well established among academics and practitioners. Although the starting point was limited to the micro-scale, the spatiality of such links gained importance over time. Initially the focus was on interfirm links, usually representing buyer–supplier relationships. The dominant economic school of thought, grouped here for simplicity reasons under the term mainstream economics, incorporated space as “transaction cost.” Thus, proximity, mostly in terms of geographic proximity among actors of local production systems, can result in transaction cost reduction. Such a proximity effect decreases the locational disadvantage of peripheral, but still agglomerated regions when compared with the core urban areas, which represent the nodal points of the Fordist production system. A similar effect can be generated by the so-called Marshallian externalities. According to Alfred Marshall's argumentation on an *industrial atmosphere*, small and agglomerated firms in *industrial districts* can benefit by scale economies and the resulted cost decrease in human capital training, information and knowledge sharing, and the collective use of infrastructure ([Marshall 1927](#)).

However, the introduction of the *innovative milieu* concept by the French group GREMI (Groupement de Recherche European sur les Milieux Innovateurs) turned the interest to more dynamic and territorially embedded conceptualizations of the innovation production. Indeed, proximity is in the center of the discussion around innovation production, not only because it decreases transaction costs, usually expressed in the form of transportation cost, but mostly because it eases information exchange usually through frequent face-to-face interactions and cultural similarity.

In more detail, there are two main characteristics of an innovative milieu. Firstly, a milieu, in order to enhance the innovation production, needs to be characterized by a *collective learning process*. Such an attribute improves local creativity, technological development, and adaptation and, in total, supports the innovation production. Secondly and in the same spirit, a local milieu reduces uncertainty as it enables a better understanding of firms' decisions given the easiness it provides in interpreting technological transformation (Camagni 1991).

At the same time, Porter (1990) introduced the notion of *clusters* and defined it as a group of interconnected companies from more than one industry and the related institutions, which are all located in geographic proximity. The notion of the industrial district differs from the regional cluster as the latter mostly consists of firms from the same or related industries as well as supporting institutions. Although he did not explicitly focus on networks, the distinctive point of his conceptual proposition was the interconnectedness of regional actors and how such network-like structures can benefit firms. The mechanism behind this argument is summed up by the following: once a specialized industrial cluster is formed, then demand for suppliers and other supporting services is generated. This demand acts as an incentive for suppliers and service producer firms to locate within the cluster to take advantage of the emerging economies of scale and the lower transaction and transportation costs. Consequently, dense transaction linkages should be expected within a cluster (Bathelt et al. 2004).

In total, various closely or less closely related concepts have been introduced in the new regionalism tradition including, apart from the innovative milieu and Porter's cluster notion, concepts such as local knowledge communities and learning regions. Even when the discussion moves to older conceptualizations such as Marshallian industrial districts, an important element for the success both of the firm and the district is the density of social networks among organizations. In both conceptualizations of local and regional networks, economic and social relations play a key role in innovation facilitation.

Therefore, given the wide acceptance of the above conceptualization, the spatiality of networks between actors came to the forefront of academic research, and in addition, such a network framework was further expanded to include non-firm actors. Examples include links between firms and other noncommercial organizations such as universities, research institutes, technology transfer agencies, and regional and governance bodies. Despite the different conceptualizations found in the literature, there are some common characteristics in respect to the mechanisms related with spatial innovation network formation (Giuliani 2011):

- Market relationships or *traded interdependencies* using Storper's (1997) terminology. Such networks are formed utilizing market mechanisms, and examples include user-producer links, spin-off companies, and highly skilled human capital mobility. Spatial proximity is an important facilitator of knowledge and most importantly tacit knowledge exchange.
- Social ties or *untraded interdependencies* (Storper 1997). More often than not, interpersonal networks provide the necessary basis to build market linkages. Examples include the importance of social relationships in the formation of

Italian industrial districts as well as their role in the success of Silicon Valley (Saxenian 1994).

- Policy-driven ties, which are the outcome of specifically designed (usually regional) policies, the main objective of which is to enhance the density of local interorganizational ties. The policy-related discussion around RIS tends to include such measures.

Despite the growing interest on networks, the relevant literature is still missing a universally accepted definition. As a summary of the above discussion, Tijssen's (1998, p. 792) definition is adopted here. According to him, a network is

an evolving mutual dependency system based on resource relationships in which their systemic character is the outcome of interactions, processes, procedures and institutionalization. Activities within such a network involve the creation, combination, exchange, transformation, absorption and exploitation of resources within a wide range of formal and informal relationships.

The discussion around networks is usually related with interfirm interactions. Firms benefit by participating in such regional innovation networks. Firstly, firms profit from the informal exchange of resources including knowledge and know-how, which is difficult to be facilitated by formal market mechanisms. Secondly, owing to an already adopted network perspective from a firm, the creation of new links is easier to take place, indicating a cumulative advantage. And thirdly, such networks carry trust and reciprocity – qualities that were usually excluded from mainstream economics, which enable the preservation of ties over long periods of time (Giuliani 2011).

Different typologies can be built for such networks. From a managerial perspective, innovation networks can be identified as the summation of interfirm linkages. Fischer (2006) distinguishes five different network types. *Supplier networks*, apart from intermediate goods purchase, reflect mostly subcontracting relationships between firms. The latter differs from the former as subcontracting relationships are ruled by formal contracts which define the order specifications, instead of simply obtaining final products from an upstream supplier. *Customer networks* represent the effort spent by companies to gain feedback from their customers to better customize their products and services. At the same time, the opposite direction of such a relationship is also important as customers – in this case firms – need to obtain information for new products from their suppliers. *Technological cooperations* is the third type of interfirm networks. The main objective of such networks is to share scientific knowledge and the outcome of R&D processes as well as to facilitate joint technology production and process development. *R&D cooperation networks* represent a well-established framework for interaction between firms and other actors such as universities and research centers, the main characteristics of which are fundamental or applied research. What used to be depicted as informal networks between firms and universities, nowadays, is mostly reflected in formal agreements between such actors. Finally, *production networks and strategic alliances* represent interfirm collaboration agreements for joint production. The main motivation for such agreements is the achievement of economies of scale and also surplus or scarcity in production capacity. While the first two types

of interfirm networks can be characterized as vertical networks, the last three represent horizontal collaborations (Fischer 2006).

From a functional point of view, interfirm networks can fulfill three different corporative needs: (a) problem solving by assistance networks; (b) lack of information such as whom to contact for specific reasons through information networks; and (c) entrepreneurship and product development (Mønsted 1993). The above networks are vastly based on interpersonal relationships and trust found in local communities. Although these networks are mostly interfirm networks, the driving force for such structures is interpersonal relations. In general, such regional innovation networks are always based on preexisting personal social networks which carry the necessary trust (Lechner and Dowling 1999). Thus, to the extent that such networks are related with regional innovation production as a necessary condition for firms to innovate and thrive, then trust, which is a necessary condition for the creation and maintenance of such interpersonal links, can be approached as an important mechanism for regional innovation processes. However, trust is a complex notion and can be approached as a socially embedded notion which is based on friendship, kinship, and repeated interaction (Boschma and Frenken 2010). Although such relationships are primarily social, they also carry information about potential partners, and because of this attribute, they increase the ability of organizations to get involved with innovation networks.

What also needs to be highlighted here is the interrelation between regional networks and regions per se. Regions do not only contain and shape network formation, but at the same time social interaction among actors in regional networks also affect its geography (Storper and Walker 1989). The above interrelation can be further intensified by the evolutionary nature of regional networks. For instance, the existence of large firms in the development stage of a region might prevent the development of extensive regional innovation networks or even an industrial district (Lechner and Dowling 1999). The explanation behind this influential role of large companies on regional networks is the power that large firms hold in terms of bargaining power against their supplies (Porter 1980). The utilization of such power has negative effects in developing interpersonal to innovation networks. A well-known example is the comparative discussion between Silicon Valley and Route 128. While in the former, the existence of one large firm, that is, Fairchild, supported the establishment of innovation networks, and the presence of multiple large firms in Route 128 prevented the creation of innovation links (Saxenian 1994).

26.3 Knowledge Networks in a Knowledge-Based Economy

Knowledge exchange is the main incentive behind the formation of networks. Knowledge production does not solely depend on isolated for-profit firms and nonprofit institutions. Knowledge comes as the outcome of unconstraint exchange of information between a plethora of actors organized formally or informally in systemic ways creating networks at different scales. Such networks tend to be more

and more inter-sectoral, interorganizational spanning over a variety of actors from firms to universities and also international (Autant-Bernard et al. 2007).

The importance of knowledge in the current economic framework is depicted in the discussion around the *knowledge economy*. Knowledge is directly linked to information because “knowledge is more than information as information is more than simply data” (Malecki and Moriset 2008, p. 29). The relation between these notions is hierarchical as one step higher in the hierarchy reveals a higher level of sophistication, codification, and consequently value. Leydesdorff (2006, 17; original emphasis) further explains the notion of knowledge and distinguishes it from information:

Knowledge enables us to codify the meaning of information. Information can be more or less meaningful given a perspective. However, meaning is provided from a system's perspective and with hindsight. Providing meaning to an uncertainty [...] can be considered as a first codification. Knowledge enables us to discard some meanings and retain some others in a second layer of codifications. In other words, knowledge can be considered as a meaning which makes a difference. Knowledge itself can also be codified, and codified knowledge can, for example, be commercialized.

This last point is the key characteristic of the knowledge economy: knowledge, as a commercialized entity, has become a production factor, in advance of capital and labor (Drucker 1998). According to the OECD's (1996, p. 7) definition, knowledge-based economies are economies “which are directly based on the production, distribution and use of knowledge and information.”

The notion of knowledge is tied with the notion of learning. The latter, as a collective ability of a society or a locale, appears to be central in the development process (Lundvall 1992). Advances in information and communication technologies (ICTs) resulted in the acceleration of codification and digitization of “codifiable” knowledge and thus in improvements in knowledge accessibility. These drastic changes in knowledge codification process resulted in the transformation of knowledge into a market commodity. The part of knowledge which is not codifiable is identified as tacit knowledge and is embodied in practices, people, and networks (Maignan et al. 2003).

Given the current knowledge economy framework, the importance of innovation and knowledge networks becomes more evident both at the micro- (firm) and meso-(city) level (for a city-level discussion around the knowledge economy, see Geenhuizen and Nijkamp 2012). *Knowledge spillovers*, which are defined as the positive externalities that a firm benefits from in terms of knowledge as a result of the environment it operates within (for a discussion, see Capello and Faggian 2005), are essential elements for knowledge generation and most importantly for the innovation creation process. Knowledge spillovers can be understood as incentives for the formation of formal and informal networks. Simply put, knowledge creation and innovation production are not just the product of one actor, but on the contrary are facilitated by spillover effects and efforts taking place outside the individual actor. Network structures support this process as they provide the necessary platform for utilizing such spillovers (Fischer 2006). Knowledge and innovation creation can be approached as interactive processes in which actors, which possess

different types of knowledge, interact together in order to overcome technical organizational, commercial, or intellectual problems (Bathelt et al. 2004).

Of equal importance in understanding the role of networks in innovation process is the distinction between tacit and codified knowledge and how these different types of knowledge are tied to specific localities. The distinctive point between them is the easiness to be transferred. For instance, codified knowledge can be digitized and transferred through ICTs over long distances without the need for intensive interpersonal interaction. Especially nowadays, owing to the digital revolution and the pervasive character of ICTs, codified knowledge can be very easily transferred or even downloaded via digital networks. Therefore, the spatial ties of this type of knowledge are loose. However, this argument applies less for tacit knowledge. The latter is characterized by a higher level of sophistication and complexity which does not allow for its codification (It needs to be highlighted here that the borderline between codification and non-codification is not fixed and may change over time). As a consequence, tacit knowledge transfer is heavily based on interpersonal interactions. Therefore, tacit knowledge has a higher degree of local embeddedness as proximity is a crucial factor for transmission of such knowledge. The role that ICTs play in supporting distant, face-to-face interactions via teleconference applications needs to be highlighted here. Face-to-face communication can be divided into two components: the *conversation* and the *handshake*, with the former being the “metaphor for simultaneous real-time interactive visual and oral messages” while the latter for the physical co-presence (Leamer and Storper 2001, p. 4). ICTs can only lower the cost of the conversation component of the face-to-face communication, which enables up to a certain extent the transmission of tacit knowledge via global networks.

The above argumentation can be crudely summarized on the preposition that the more codified knowledge is, the less spatial dependent is. On the other hand, the more tacit characteristics knowledge has, the more its transfer is based on spatial proximity between the involved actors. This spatial distinction of knowledge and the knowledge transfer process also reflects the debate between global connectivity and local network intensity. The discussion in this chapter has been mostly focused on the latter, which refers to Marshallian externalities, clusters, and innovative milieu. All of these notions highlight the value of local networks in innovation generation and growth. What has not been yet discussed in this chapter is the value of global links in achieving the above objectives. Such global *pipelines* can be understood both as open channels and more closed conduits. The former approaches interorganizational links as open systems that diffuse knowledge to all the loosely connected actors in a way that facilitates knowledge spillovers. This type of linkages can be understood as weak links. The latter type of links functions in a more restricted way so that knowledge only flows among these connected actors that are part of the alliance. Such strong links are used to protect sensitive issues such as intellectual property rights. Based on this, it can be said that access to knowledge is not only the result of interactions among collocated actors in local or regional networks, but it can also be the outcome of partnership and linkages at an interregional or international scale (Owen-Smith and Powell 2004).

The interplay between local and global linkages can also be seen as the outcome of the transformation that our society is going through due to the extensive use of ICTs. From an urban perspective, the underlying new techno-economic paradigm is related with drastic social changes. The starting point for understanding these changes is the seminal work of Castells on the *network society* (Castells 1996), where he illustrated the emergence of a new spatial form due to the structural transformation that society is undergoing because of the extensive use of ICTs. He identified this new spatial form as *the space of flows*, and he defined it as the “managerial organization of time-sharing social practices that work through flows” (Castells 1996, p. 442). Such flows are “purposeful, repetitive, programmable, sequences of exchange and interaction” between detached socioeconomic actors (Castells 1996, p. 442). Castells presented this new spatial form as a three layer system. The first layer consists of the technical network infrastructure, upon which the flows of Castells’ network society are transported. Examples of such infrastructure include the digital infrastructure upon which the Internet function is based as well as aviation networks which are responsible for transfer of people between places. Most importantly from the innovation network viewpoint, the second layer refers to the nodes and the hubs of the space of flows. These are the real places with “well-defined social, cultural, physical, and functional characteristics” (Castells 1996, p. 443). These places – cities in reality – are interlinked through the first – infrastructural – layer of the space of flows upon which real flows, such as knowledge, are transported. From this perspective, global pipelines can be understood as part of the space of flows. Lastly, the third layer of the space of flows refers to “the dominant managerial elites” and analyzes the spatial organization of these privileged social groups, which are increasingly located in isolated communities, but at the same time in highly connected places (Castells 1996, p. 433).

What would be a mistake here is to approach the above discussion on the scale of interorganizational linkages as preference toward local or global links. The argument that innovation cannot occur in isolation is only valid from a multi-scalar perspective in the contemporary world economy. It is well established nowadays that local economies are dependent upon global corporate processes. Thus, localities cannot exist anymore as local and regional economies only linked with the global economic system via trade flows, following a Marshallian logic. From a policy point of view, strategies to support local economies using a Marshallian framework targeting only the intensification of local links ignore global interdependencies, and their success is anything but given (Amin and Thrift 1992). Such local strategies should be multi-scalar in nature and promote cross-fertilization between global and local links.

26.4 Innovation Networks and Different Types of Proximity

The discussion in the previous section took place on two axes. On the one axis, different knowledge types were analyzed. The focus was on tacit and codified knowledge and how these “distinct” types of knowledge flow between

organizations. On the other axis, the scale and the spatiality of interorganizational linkages, which facilitate these flows, were examined. As it was explicated, the range of such links varies from local to global links. Despite the importance of the above approach in understanding the mechanisms behind the formation of knowledge and innovation interorganizational links, it is still a simplification to assume an absolute correspondence between tacit knowledge and geographic proximity, on the one hand, and codified knowledge and long distance links, on the other hand. Firstly, the separation of both knowledge types is not always clear. Secondly, face-to-face interactions and geographic proximity are not the exclusive facilitators of tacit knowledge flows. For instance, ICTs, as mentioned before, can also play a role as such a facilitator. Most importantly though, it is important for innovative firms to establish links with nonlocal partners to obtain knowledge and ideas which are not accessible locally (Torre 2008).

In order to shed more light in this complex relation, this section takes a relational turn and introduces other non-geographic components of proximity. Thus, apart from the geographical proximity denoted by Euclidean distance, other relational proximities, including cognitive, organizational, and institutional proximity, will also be assessed. These proximities are defined in a *relational space* which can be defined as “the set of all relationships – market relationships, power relationships and cooperation – established between firms, institutions and people that stem from a strong sense of belonging and a highly developed capacity of cooperation typical of culturally similar people and institutions” (Capello and Faggian 2005, p. 78). The starting point for defining the different dimensions of proximity lies in the French school of proximity. The main objective of this group of industrial economists was to endogenize space in economic analysis and, more specifically, to incorporate space and other territorial proximity elements in a research framework, which aims to better understand the dynamics of innovation (for a review of the French school of proximity, see Torre 2008). A second development in further decomposing and analyzing the different components of proximity was studies related to innovation and territorial learning in the broader framework of evolutionary economic geography. In recent years, we have experienced an increased interest in factors which explain how firms and regions interact as part of a “collective learning process,” since learning and knowledge creation are an essential component of the firms’ and regions’ competitive advantage. The notion of proximity and its different components is juxtaposed with ideas about knowledge transfer and creation, tacit knowledge, and learning regions (Boschma 2005).

The common basis of these approaches is the importance of non-geographic types of proximity in innovation creation. Starting from the French school of proximity, two different types of proximity can be identified: geographical and organized (Torre 2008). The former type is more straightforward and usually represents physical distance and collocation. Nonetheless, different conceptualizations of geographic proximity could also be utilized, as physical proximity might be affected not only by Euclidean distance but also by the transportation cost between two places and their accessibility. In addition, the temporal continuity of

geographic proximity is also important. As Torre (2008) suggests, geographic proximity and face-to-face interactions are necessary only during specific stages of the innovation process.

Unlike physical proximity, organized proximity is a relational notion and refers to the ability of an organization to enhance interaction between its members. The main point behind this concept is that members of the same organization will interact together more easily than actors outside the organization. This is based on two different logics (Torre 2008): (a) adherence logic, according to which actors, who are close in organizational terms, such as a firm and network, are part of the same relational space, and (b) similarity logic, according to which the organizationally close actors tend to be alike. In a nutshell, while geographical proximity reflects separation in space regarding physical distance, organized proximity is considered as the overall framework in which different actors interact. In the same vein, Boschma defined organizational proximity “as the extent to which relations are shared in an organizational arrangement, either within or between organizations” (Boschma 2005, p. 65).

Building upon the French school of proximity findings, Boschma (2005) approached proximity as a five-dimension notion. Cognitive proximity is the point of departure for his conceptual framework and is defined as the level of similarity of the knowledge base of different organizations. Organizations collaborate and form links and networks using as criteria the knowledge background of the potential partners, as people and organizations, which share the same knowledge background and expertise, may learn from each other. The complexity of the learning process is reflected in the nonlinear effect of cognitive proximity in learning. In order for knowledge to be transferred, gained, or created, there is a need for optimal cognitive proximity, since too high cognitive proximity will eliminate any novelty from the interaction, while, vice versa, too high cognitive distance will result in communication difficulties (for a detailed discussion, see Boschma 2005).

Despite the effort spent in the relevant literature, cognitive proximity is still a rather fuzzy concept, and it is difficult to quantify. Strong links can be identified between cognitive proximity and technological similarity. While some authors distinguish these two, more often than not, these notions are used interchangeably in an empirical context. While cognitive proximity represents the similarity of the knowledge bases of two organizations or regions, technological proximity reflects the similarity between the technological knowledge among economic actors (Dangelico et al. 2010).

In addition, institutional proximity is also proposed as another proximity dimension. Following North’s (1990) definition, institutions are the amalgamation of formal rules and informal constraints including behavioral and social standards, while organizations can be approached as a group of agents performing the same activity. Put simply, organizations define agents’ practices and strategies in the overall context provided by the institutional ecosystem in which they are positioned (Kirat and Lung 1999). Therefore, one would expect that collocation in the same institutional environment would result in increased interaction.

The analytical value of the multidimensional approach of proximity is now well established in the relevant literature as a tool for understanding structurally interorganizational innovation networks. Apart from the above discussed proximity dimensions, different proximity components can be introduced such as social or linguistic proximity. Most importantly, in a quantitative framework, a multidimensional understanding of proximity will enable researchers to compare the impact of different proximity measures in the formation of linkages. Such a framework provides researchers with the necessary tools and flexibility to model the explanatory value of different proximity dimensions in network formation. For instance, it can be claimed that the different proximity dimensions in innovation networks are in reality substitutes rather than complements (Boschma 2005).

Despite the increased complexity that the inclusion of relational proximities introduces to the discussion on interorganizational spatial innovation networks, the above proximity components reveal interesting dimensions of organizational behavior. Thus, the inclusion of these relational dimensions of proximity should not be interpreted as a diminishing factor of the importance of the spatiality of interorganizational networks. It might be the case that geography, in terms of collocation and geographical proximity, is not the (only) determinant for the innovation networks formation, but on the other hand, the spatiality of the other relational dimensions of proximity might reveal interesting geographies as well. For instance, path dependency could explain organizational proximity between geographically distant actors. Although geography could not explain this phenomenon, a study of the spatiality of such networks and the underlying relational proximities, which do not correspond with the geographical one, could be of interest per se.

26.5 Innovation Networks: Some Methodological Approaches

The centrality of spatial innovation networks in economic geography and regional science can be depicted on the numerous different methodologies adopted by researchers in their effort to understand and model innovation activity and innovation networks in space. In this section, some recent methodological advances will be presented. On the one hand, “traditional” econometric modeling is enriched with spatial econometric concepts to understand innovation in space. Although this strand of research is less associated with network structures, it still provides insights on the spatiality of the innovation process and therefore is briefly presented here. On the other hand, the network structure of innovation activity is the key focus on studies having a starting point on *network science*. Despite the fact that space and geography is not the key focus of this field, recent developments incorporate this dimension as well in the search of the innovation networks driving forces.

A starting point for the first strand of methodological approaches is the *knowledge production function* which relates regional knowledge output with R&D by industry and university research in a Cobb-Douglas framework. The original work of Zvi Griliches and Adam B. Jaffe was further expanded by the developments in the spatial econometrics (for a discussion, see Anselin et al. 1997). The latter

enabled an in-depth investigation of the spatiality of knowledge spillovers. For instance, the work of Anselin et al. (1997) confirmed the significant positive relationship between university research and innovative activity, both directly and indirectly, through its impact on private sector R&D. Most importantly, the spatiality of university research spillovers on innovations was confirmed and quantified.

Moving to models addressing the relational character of interorganizational links, gravity models, an important model class in regional science, may be used to measure the impact of different types of proximity or distance on knowledge flows. These models typically rely on three types of factors: (i) origin-specific factors that characterize the ability of origin locations to generate knowledge flows, (ii) destination-specific factors that represent the attractiveness of destination locations to absorb knowledge flows, and (iii) origin–destination factors that characterize the way spatial separation of origin from destination locations constrains or impedes the interaction (Fischer and Wang 2011). Suppose that we have a spatial system of n locations representing network nodes, then the following (lognormal) knowledge flow model may be taken as a framework for the analysis:

$$\ln kF_{ij} = \alpha_0 + \alpha_1 \ln X_i^o + \alpha_2 \ln X_j^d + \alpha_3 D_{ij} + \alpha_4 BD_{ij} + \varepsilon_{ij} \quad i, j = 1, \dots, n \quad (26.1)$$

where kF_{ij} represents the knowledge flow (e.g., measured in terms of patent citations) from origin location i ($i = 1, \dots, n$) to destination location j ($j = 1, \dots, n$). X_i^o and X_j^d are origin-specific and destination-specific factors, respectively, and α_1 and α_2 denote the associated coefficients. D_{ij} is a continuous distance (proximity) measure, for example, geodetic distance or travel time from i to j , and BD_{ij} is a binary distance measure representing, for example, institutional proximity between the locations, with the corresponding distance sensitivity coefficients α_3 and α_4 . α_0 is the constant, and ε_{ij} is the error term, generally assumed to be identically and independently distributed. For model estimation issues, and econometric extensions of the lognormal gravity model to account for spatial or network dependence in flow data, see Fischer and Wang (2011, p. 47–70).

Finally, network science is shaping the latest developments in modeling the structure and the evolution of innovation networks, also from a spatial perspective. From a descriptive point of view, network analysis provides a plethora of metrics which can assist researchers to understand the topology of innovation networks. Examples include different *centrality* indicators, which depict the position of an actor and the roles it performs in the overall network, the *clustering coefficient*, which indicates the tendency of a network to create clusters of dense internal connections, and the *average path length*, which is a measure of network distance. From a modeling point of view, complexity science provides a plethora of tools to model innovation networks from a structural and evolutionary perspective. The distinctive point of this strand of research is that instead of adopting an explanatory modeling strategy as reflected in fitting regression lines in observed data (e.g., the gravitation family of models discussed above), stochastic models are utilized to

understand the underlying mechanisms of the network formation and to simulate the evolution of the (observed) network. A wide range of such modeling applications – from agent-based models to statistical physics and social network analysis – can be found in the book edited by Pyka and Scharnhorst (2009).

To sum up, spatial innovation networks are curved by a two-level complexity: a network level complexity which reflects the topological characteristics of k interacting actors and a spatial level complexity which represents the peculiarities of geographical space where innovation interactions are embedded. Despite the difficulty to model these mechanisms, studies focusing on such issues are currently at the forefront of regional science. Nonetheless, the incorporation of both level complexities is still a challenging task despite the developments in modeling presented in this section.

26.6 Conclusions

The objective of this chapter was to highlight the importance of networks in the innovation process. From the above discussion, it became apparent that innovation does not occur in isolation and that interorganizational networks facilitate innovation creation. Such networks are usually based on interpersonal relations characterized by a high level of trust. Moreover, these social networks are embedded in places. This reflects the importance of the spatiality of such networks. In addition, such innovation networks can be understood as knowledge networks and knowledge externality catalysts. Actors interact to gain knowledge, a crucial production element of the knowledge economy. Broadly speaking, two different types of knowledge and two scales of interaction can be identified: tacit and codified knowledge and local and global interaction. Although there is no absolute correspondence, it can be said that tacit knowledge is facilitated by geographic proximity, while codified knowledge can be easily accessed remotely. Counterarguments to the above statement are the increased use of teleconferencing applications via desktop computers and the necessity for firms to establish long-haul links to gain knowledge which is not available locally. These arguments advocate toward the adoption of a relative proximity perspective. The latter enables researchers to understand the determinants of innovation activity and networking not only in the Cartesian but also in relative space. Finally, effort was spent to approach innovation networks from a systemic perspective. Innovation networks are vital parts of multi-scalar systems of innovation, and this systemic attribute should also be reflected in attempts to model such networks as well as their dynamics.

To conclude, despite the inherent complexity for understanding and modeling such networks, especially from a spatial perspective, the research community should continue its efforts for two reasons. Firstly and from an analytical perspective, we are still lacking a generalized understanding of how actors interact together in order to innovate. From a spatial perspective, it is essential to understand the role

places perform in this process. Secondly, such analytical gains can be utilized to better design local and regional policies. In the current network society, it is important for policy makers to propose tools for increasing network intensity in a multi-scalar, targeted, and efficient way.

References

- Amin A, Thrift N (1992) Neo-Marshallian nodes in global networks. *Int J Urban Reg Res* 16(4):571–587
- Anselin L, Varga A, Acs ZJ (1997) Local geographic spillovers between university research and high technology innovations. *J Urban Econ* 42(3):422–448
- Autant-Bernard C, Mairesse J, Massard N (2007) Spatial knowledge diffusion through collaborative networks. *Pap Reg Sci* 86(3):341–350
- Bathelt H, Malmberg A, Maskell P (2004) Clusters and knowledge: local buzz, global pipelines and the process of knowledge creation. *Prog Hum Geog* 28(1):31–56
- Boschma R (2005) Proximity and innovation: a critical assessment. *Reg Stud* 39(1):61–74
- Boschma R, Frenken K (2010) The spatial evolution of innovation networks. A proximity perspective. In: Boschma R, Martin R (eds) *The handbook of evolutionary economic geography*. Edward Elgar, Cheltenham
- Camagni R (1991) Innovation networks: spatial perspectives. Belhaven Press, London/New York
- Capello R, Faggian A (2005) Collective learning and relational capital in local innovation processes. *Reg Stud* 39(1):75–87
- Castells M (1996) *The rise of the network society*. Blackwell, Oxford
- Cooke P (1992) Regional innovation systems: competitive regulation in the new Europe. *Geoforum* 23(3):365–382
- Cooke P, Morgan K (1993) The network paradigm: new departures in corporate and regional development. *Environ Plann D: Soc Space* 11(5):543–564
- Dangelico RM, Garavelli AC, Petruzzelli AM (2010) A system dynamics model to analyze technology districts' evolution in a knowledge-based perspective. *Technovation* 30(2):142–153
- Drucker PF (1998) From capitalism to knowledge society. In: Neef D (ed) *The knowledge economy*. Butterworth-Heinemann, Woburn, pp 15–34
- Edquist C (1997) Systems of innovation approaches—their emergence and characteristics. In: Edquist C (ed) *Systems of innovation: technologies, institutions and organizations*. Pinter, London, pp 1–35
- Fischer MM (2006) *Innovation, networks and knowledge spillovers*. Springer, Berlin/Heidelberg
- Fischer MM, Wang J (2011) *Spatial data analysis. Models, methods and techniques*. Springer, Heidelberg/Dordrecht/London/New York
- Giuliani E (2011) Networks of innovation. In: Cooke P, Asheim B, Boschma R, Martin R, Schwartz D, Tödtling F (eds) *Handbook of regional innovation and growth*. Edward Elgar, Glos/Northampton, pp 155–166
- Kirat T, Lung Y (1999) Innovation and proximity: territories as loci of collective learning processes. *Eur Urban Reg Stud* 6(1):27–38
- Leamer EE, Storper M (2001) The economic geography of the internet age. *J Int Bus Stud* 32(4):641–665
- Lechner C, Dowling M (1999) The evolution of industrial districts and regional networks: the case of the biotechnology region Munich/Martinsried. *J Manage Gov* 3(4):309–338
- Leydesdorff L (2006) The knowledge-based economy: modelled, measured, simulated. Universal Publishers, Boca Raton, FL
- Lundvall B-A (ed) (1992) *National innovation systems: towards a theory of innovation and interactive learning*. Pinter, London

- Maignan C, Pinelli D, Ottaviano GIP (2003) ICT, clusters and regional cohesion: a summary of theoretical and empirical research. <http://www.ssrn.com/abstract=438507>. Accessed 13 Jul 2011
- Malecki EJ, Moriset B (2008) The digital economy. Routledge, New York
- Marshall A (1927) Industry and trade. a study of industrial technique and business organization; and their influences on the conditions of various classes and nations, 3rd edn. Macmillan, London
- Mønsted M (1993) Regional network processes: networks for the service sector or development of entrepreneurs? In: Karlsson C, Johannesson B, Storey D (eds) Small business dynamics. Routledge, London, pp 204–222
- Geenhuizen MS, van Nijkamp P (eds) (2012) Creative knowledge cities: myths, visions and realities. Edward Elgar, Glos/Northampton
- North DC (1990) Institutions, institutional change and economic performance. Cambridge University Press, Cambridge
- OECD (1996) The knowledge-based economy. OECD, Paris
- OECD and Eurostat (2005) Oslo manual: guidelines for collecting and interpreting innovation data. OCED and Eurostat, Paris
- Owen-Smith J, Powell WW (2004) Knowledge networks as channels and conduits: the effects of spillovers in the Boston biotechnology community. *Organ Sci* 15(1):5–21
- Porter M (1980) Competitive strategy. Free Press, New York
- Porter M (1990) The competitive advantage of nations. Free Press, London
- Pyka A, Scharnhorst A (eds) (2009) Innovation networks: new approaches in modelling and analyzing. Understanding Complex Systems. Springer, Berlin/Heidelberg
- Saxenian A (1994) Regional advantage: culture and competition in Silicon Valley and Route 128. Harvard University Press, Cambridge, MA
- Storper M (1997) The regional world. The Guilford Press, New York
- Storper M, Walker R (1989) The capitalist imperative: territory, technology, and industrial growth. Blackwell, Oxford
- Tijssen RJW (1998) Quantitative assessment of large heterogeneous R&D networks: the case of process engineering in the Netherlands. *Res Policy* 7–8(26):791–809
- Torre A (2008) On the role played by temporary geographical proximity in knowledge transmission. *Reg Stud* 42(6):869–889

Section IV

New Economic Geography and Evolutionary Economic Geography

city goods evolution
agents reg.
growth productivity rent
industry technological
development distance
workers transport
knowledge clusters
revolutionary technological
firm competition
geography consumers
consumers clusters
market countries
agglomeration change
firms NEG
demand land trade
social proximity
spatial activities wages
local proximity
innovation production
costs

Roberta Capello

Contents

27.1	Introduction	508
27.2	The Location of Activities in Space: Land Rent Formation	510
27.2.1	Accessibility and Transportation Costs	510
27.2.2	The Location of Agricultural Activities: The von Thünen Model	511
27.2.3	The Location of Activities in a City: A Partial Spatial Equilibrium Approach	514
27.2.4	The Location of Activities in a City: A General Spatial Equilibrium Approach	515
27.2.5	Critical Remarks	519
27.3	The Location of Industrial Activities in Space: Weber's Model	520
27.3.1	Transportation Costs and Agglomeration Economies	520
27.3.2	Weber's Model	521
27.3.3	Criticisms of the Model	523
27.3.4	Strength of the Model	524
27.4	Conclusions	525
	References	525

Abstract

Within location theory, classical models are typical abstract and formalized models, in which the main reasoning behind location choice of firms is driven by the *minimization of transportation costs* to achieve natural and intermediate production resources, and markets for final goods that are territorially dispersed. Classical models are similar in the question they want to reply to: what economic logic explains the location choices of firms in space? This topic is an important one.

R. Capello

Department Architecture, Built Environment and Construction Engineering A.B.C., Politecnico di Milano, Milan, Italy
e-mail: roberta.capello@polimi.it

Although in terms of time and financial resources, the performance of transport and communication has improved enormously, many economic activities have not become footloose to the extent expressed by the “death of distance.” Their location choice still remains anchored to a balance between a physical location generating economic advantages – in the form of agglomeration economies – and transport costs to intermediate or final markets, as explained by these models.

27.1 Introduction

Economic activity arises, grows, and develops in space. Firms, and economic actors in general, choose their locations in the same way as they choose their production factors and their technology. Location theory aims at explaining the economic rationale behind the choice of a firm to locate in a specific point in space and thus at interpreting the allocation of different portions of territory among different types of production, the division of a spatial market among producers, and the functional distribution of activities in space.

This topic is an important one; although in terms of time and financial resources, the performance of transport and communication has improved enormously, many economic activities have not become footloose to the extent expressed by the “death of distance” (Cairncross 1997). The location choice of firms still remains anchored to a balance between advantages that a physical location generates for firms – in the form of agglomeration economies – and transport costs, keeping the role of space in modern economies an important issue.

Within location theory, classical models are typical abstract and formalized models, in which the main reasoning behind location choice of firms is driven by the *minimization of transportation costs* to achieve natural and intermediate production resources, and markets for final goods that are territorially dispersed. Localization theory defines “transportation costs” as all the forms of spatial friction that give greater attractiveness to a location which reduces the distance between two points in space (e.g., production site and the final market, place of residence and the workplace, the raw materials market, and the production site). Transportation costs are essential to location theory, in general, and to industrial location choice models, in particular, for a very simple reason. The value of two production resources located in two different points in space can be directly compared only if the physical distance between the two resources is discounted. Transportation costs (i.e., the cost of movement of such inputs in space) represent the discount rate of *space*, as interest rates represent the discount rate of values in *time* (Isard 1956).

It is important to notice that in all location theory, transportation costs have a wider meaning than that of the economic cost of shipping goods (the pure cost of transporting and distributing them); they refer more in general to the opportunity cost represented by the time taken to cover the distance which could instead be put to other uses, by the psychological cost of the journey, by the cost and difficulty of communication over distances, and by the risk of failing to acquire vital information.

Transportation costs are even comprised in the concept of agglomeration economies, which denote all economic advantages accruing to firms from concentrated location close to other firms, namely, reduced production costs due to large plant size, the presence of advanced and specialized services, the availability of fixed social capital (e.g., infrastructures), the presence of skilled labor and of managerial expertise, and the presence of a broad and specialized intermediate goods market. If transportation costs were nil, there would be no reason to concentrate activities, because doing so would not produce any economic advantage. In this sense, agglomeration economies are “proximity economies”: they are, that is to say, advantages which arise from the interaction (often involuntary) among economic agents made possible by the lower amount of spatial friction in concentrated locations.

The logic behind classical location models is therefore the minimization of transportation costs between a production area and the market where goods are sold (von Thünen model), between place of residence and the workplace (Alonso model), and among production site, raw materials markets, and the final goods market, which together define a minimum transportation cost directly compared against agglomeration advantages (Weber model).

Classical models are similar in the question they want to reply to: what economic logic explains the location choices of firms? Their richness comes from what they, directly and indirectly, are able to explain. In replying to the question on where activities would be willing to locate their production in space, the von Thünen's model indirectly interprets the formation of a land rent at different distances from a market place in a general equilibrium setting. Within the same framework of assumptions of von Thünen's model, in a partial equilibrium approach, Alonso is able to explain the location choice of a new incoming firm in a city; moreover, in a general spatial equilibrium framework, Alonso is able to interpret the allocation of alternative types of activities in the urban space. In replying to the same conceptual question, i.e., “where do activities locate their production?” Weber is able to identify a location that reflects the two economic forces that organize activities in space and that push the location process in opposite directions: from one side, transportation costs, that – in conditions of perfect competition, perfectly mobile production factors, fixed raw materials and demand perfectly distributed across the territory – induce dispersion and, from the other, agglomeration economies that call for spatial concentration of production.

In what follows, the two main classical location theory models are presented. Firstly, the von Thünen-Alonso model is presented based on the assumption that the production site assumes a spatial dimension and extends across a territory, while the consumption site (the market) is punctiform. In this way, the model defines a “production area,” meaning by this the physical space (the land) occupied by an individual economic activity. Secondly, the Weber model is taken into account based on the assumptions that both the production and consumption sites are punctiform and that the minimization of the distance between them drives the identification of the minimum transportation cost location, directly compared

against a location generating agglomeration advantages. The models interpreting the identification of market areas, assuming that production develops at specific points in space and it supplies geographically dispersed markets, are left to the next chapter.

27.2 The Location of Activities in Space: Land Rent Formation

27.2.1 Accessibility and Transportation Costs

In the theories examined in this section, location choices are dictated by a specific principle of spatial organization of activity, namely, “accessibility,” and in particular accessibility to a market or a “center.” For firms, high accessibility means that they have easy access to broad and diversified markets for final goods and production factors, to information, and to the hubs of international infrastructures. For people, accessibility to a “central business district” and therefore to jobs means that their commuting costs are minimal, while at the same time, they enjoy easy access to a wide range of recreational services restricted to specific locations (e.g., theaters, museums, libraries) and proximity to specific services (e.g., universities), without having to pay the cost of long-distance travel.

High demand for accessibility to central areas triggers competition between industrial and residential activities for locations closer to the market or, more generally, closer to the hypothetical central business district (the city center).

The location choice models described in this section have an important feature in common: the *cost of land or land rent*. Assuming the existence of a single central business district, owing to high demand for central locations with their minimum transportation costs, land closer to the center costs more, a condition accentuated by the total rigidity, at least in the short-to-medium period, of the urban land supply. The models resolve the competition among activities on the basis of a strict economic principle: *firms able to locate in more central areas are those able to pay higher rents for those areas*.

These models envisage one specific factor organizing activity in space: land rent, this being the sole principle which explains location choices by all activities, whether agricultural, productive, or residential.

The strength of these models is the elegant and irrefutable logic with which they account for the distribution of productive, agricultural, and residential activities in a geographic space from which they eliminate every differentiating effect except for physical distance from the center. Given their assumptions on the structure of demand and supply in space, these models are particularly well suited to analysis of the location of industrial and residential activities in urban space. In an urban environment, in fact, it is easy to hypothesize the existence of a single business district (a city center) which, for firms, performs the function of collecting, distributing, and exporting the city’s products and, for households, is the place where jobs are available. These models are able to establish where an individual firm will locate.

The first model analyzing the spatial distribution of alternative production activities was developed in the early nineteenth century by Johann von Thünen. Only in the 1960s did pioneering studies by Walter Isard, Martin Beckmann, and Lowdon Wingo prepared the ground for Alonso's formulation of von Thünen's historical model applied to an urban context (Beckmann 1969; Isard 1956; Wingo 1961). The model of the monocentric city soon became a freestanding school of thought within location theory, where it was labeled "new urban economics." This corpus of theories endeavored to develop general equilibrium location models in which the main interest is no longer the decisions by individual firms. Instead, the main areas of inquiry become definition of the size and density of cities, and identification of the particular pattern of land costs at differing distances from the city that guarantees achievement of a location equilibrium for all firms in the city.

27.2.2 The Location of Agricultural Activities: The von Thünen Model

Johann Heinrich von Thünen developed the first location model based on the hypothesis of a continuous production space and a single punctiform final market (von Thünen 1826). His model has generated the entire corpus of theories on the urban location of economic activities.

Von Thünen's model is based on a set of assumptions which all subsequent theories would adopt:

- (a) There exists a uniform space where all land is equally fertile and transport infrastructures are identical in all directions (isotropic space).
- (b) There is a single center, the medieval town, where all goods are traded (i.e., there is a specific market place).
- (c) Demand is unlimited, an assumption which reflects the supply-oriented nature of the model: the location equilibrium depends solely on the conditions of supply.
- (d) The production factors are perfectly distributed in space: the allocation of land among alternative production activities does not derive from an uneven spatial distribution of the production factors.
- (e) There is a specific production function, with fixed coefficients and constant returns to scale, for each agricultural good; this assumption entails that the quantity of output obtainable from each unit of land and the unit cost of production are fixed in space.
- (f) Perfect competition exists in the agricultural goods market: farmers therefore take the prices of the goods they produce to be given.
- (g) Unit transportation costs are constant in space: the total cost of transportation depends on the distance between the production site and the town and on the volume of production. Transportation costs may vary according to the crop.

Assuming the existence of a certain number of farmers, von Thünen addresses the problem of how to determine the allocation of land among farmers working in

the area surrounding the market place. He bases his model on a concept of rent as a residual which comes from Ricardo's model and would also characterize subsequent models. Ricardo was the first to interpret rent as a differential rent (Ricardo 1971, original version 1817). In Ricardo, land rent is formed as a result of the productivity difference between agricultural lands (*decay of productive power*); if all agricultural lands had the same fertility, rent would be null. In von Thünen, lands are equal in terms of fertility but differ in terms of accessibility to the market; the difference in accessibility explains their different rent values.

In von Thünen, the price that farmers are willing to pay for land is the remainder left when transport and production costs, including a certain remuneration (profit) for the farmer, have been subtracted from revenues. In this logic, rent does not enter the formation of a good price, but is formed on the basis of the demand of products. If rent is high, it is because the demand for goods is high, and therefore the final price of goods is high; the high price of goods generates high profits and therefore high margins for rents. This means that rent is formed through the distribution of income rather than through income production. In the words of the classical economist Ricardo, “*corn is not high because rent is high, but rent is high because corn is high*” (Ricardo 1971, p. 98).

In von Thünen's model, in formal terms, if x is the quantity of a good per unit of land produced by a farmer, c the unit cost of production, p the price of the agricultural good, τ the unit cost of transportation, and d the distance to the market, then rent r is defined as

$$r(d) = (p - c - \tau d)x \quad (27.1)$$

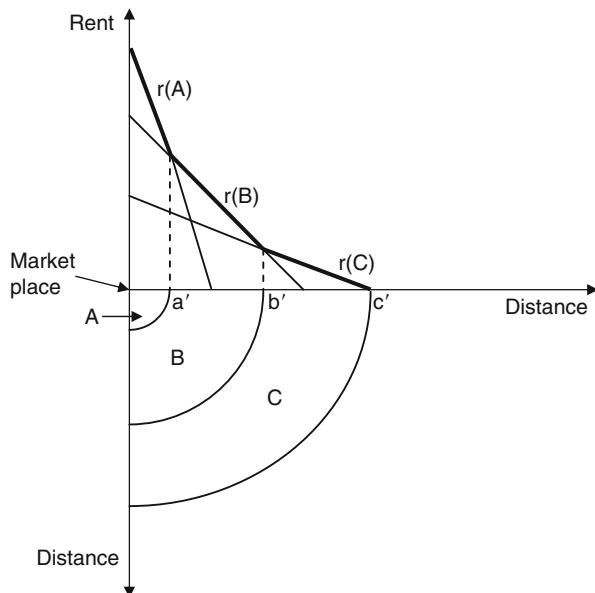
This equation states the levels of rent that farmers are willing to pay for land at different distances from the market place where goods are traded. It is represented graphically by a straight line, with slope $-\tau d$ and intercepts equal to $(p-c)x$ and $(p-c)/\tau$, respectively, denoting the maximum value of rent in the town and the maximum distance from the town, where land value is nil.

From Eq. (27.1) one can straightforwardly obtain the impact on rent due to a shift in space (e.g., of 1 km) by calculating the first derivative of rent with respect to distance:

$$\frac{dr(d)}{dd} = -\tau x \quad (27.2)$$

As Eq. (27.2) shows, the variation in rent is exactly equal to $-\tau x$: a shorter distance from the center generates a saving in total transportation costs equal to the increase in the rent required to occupy more central locations. This result is important because it has been obtained by all the models developed since von Thünen's. It states that rent is nothing other than a saving in transportation costs made possible by more central locations. From this follows the “indifference to alternative locations” condition, which is reached by an individual or a firm when a move in space costs nothing, i.e., when the saving in transportation costs

Fig. 27.1 Land allocation among three farmers: the Von Thünen model. Legend: $r(A)$ willingness to pay of farmer A, $r(B)$ willingness to pay of farmer B, $r(C)$ willingness to pay of farmer C, A area allocated to farmer A, B area allocated to farmer B, C area allocated to farmer C, a' maximum distance from the centre for which A offers a higher rent than farmers B and C, b' maximum distance from the centre for which B offers a higher rent than farmers A and C, c' maximum distance from the centre for which C offers a higher rent than farmers A and B



obtained by moving 1 km closer to the center equals the cost of the land that must be purchased to do so.

On the assumption that there are three farmers (A , B , and C), each of them producing a specific agricultural product with a differing degree of perishability, a rent supply curve can be constructed for each farmer. Because goods are perishable to differing extents, the rent supply curves assume different positions and slopes (Fig. 27.1).

The farmer who produces the most perishable good will have a productive process that uses the land in the most intensive and economically efficient way (geometrically, the highest intercept on the vertical axis, equal to $(p-c)x$), and he will be more willing to pay the rent charged for land 1 km closer to the town (geometrically, the steeper slope of the straight line, equal to $(-rx)$). As the farmers compete for the more accessible land, each unit of surface area will be allocated to the farmer willing to pay the highest rent for that land. As far as a' , the land will be allocated to farmer A , who offers the highest rent for the most central locations, from a' to b' to farmer B and from b' to c' to farmer C : the actual rent realized by the landowner from cultivation of his land is the envelope of the three rent supply curves.

This model strikes for its strong interpretative power. Assuming a homogeneous plain, with no economic activities located in this geographical space, the model is able to explain the formation of agglomeration by the simple distance from, or accessibility to, the town (expressed by transportation costs) which accounts for differences in land rent. Moreover, assuming equal fertility of land everywhere, the model departs from the classical Ricardian view that land profitability is explained by different degrees of fertility (Ricardo 1971;

orig. ed., 1817); simple distance from, or accessibility to, the town (expressed by transportation costs) accounts for differences in land rent in this model. By eliminating everything except the distance between land and the town from concrete geographic space, von Thünen defines a new type of space, namely, economic space.

27.2.3 The Location of Activities in a City: A Partial Spatial Equilibrium Approach

In the early 1960s, first William Alonso and then Richard Muth reconsidered von Thünen's model and adapted it to an urban context (Alonso 1960, 1964a, b; Muth 1968, 1969), thus paving the way for numerous subsequent studies. Alonso and Muth extended the bases of von Thünen's pioneering model, making it more specific to the urban case, but they also made it more general by abandoning the hypothesis that only transportation costs express spatial friction and the preference for more central locations.

The assumptions of Alonso's model are the same as those of von Thünen's model of agricultural activity described above. It envisages a city (no longer a plain) characterized by uniform space (homogeneous spatial distribution of the production factors) and endowed with infrastructures which cover the entire city in all directions (isotropic space). The city has a single center – the city center or business district – which is generically defined as the most attractive location for all firms. Given these assumptions, the city is analyzed along only one dimension: a radius comprising different distances from the city center to the periphery.

The base model addresses a problem similar to the one which preoccupied von Thünen. As firms compete for central locations, the model shows how the urban space is allocated among alternative types of production once the market cost of land at different distances from the center is known.

Also Alonso's model defines rent as the remainder left when the entrepreneur has subtracted production costs (including transportation costs) and a desired level of profit from the revenue obtained by selling the good. Formally, rent is expressed as

$$r(d) = (p_x - \pi - c(d))x(d) \quad (27.3)$$

where r denotes the rent, p_x the unit price of the good produced by the entrepreneur, c production costs per unit of land (including transportation costs) at distance d from the city center, π the profit per unit of land, and x the quantity of the good produced per unit of land at a distance d from the city center.

Because production costs include transportation costs, in the Alonso model, they depend on distance, as they do in von Thünen's model. However, unlike in the latter, revenues too depend on distance: a less suburban location gives greater proximity to broader markets and consequently access to higher earnings (consider the sales of a shop located in the city center compared to one in the periphery, especially if they sell luxury items).

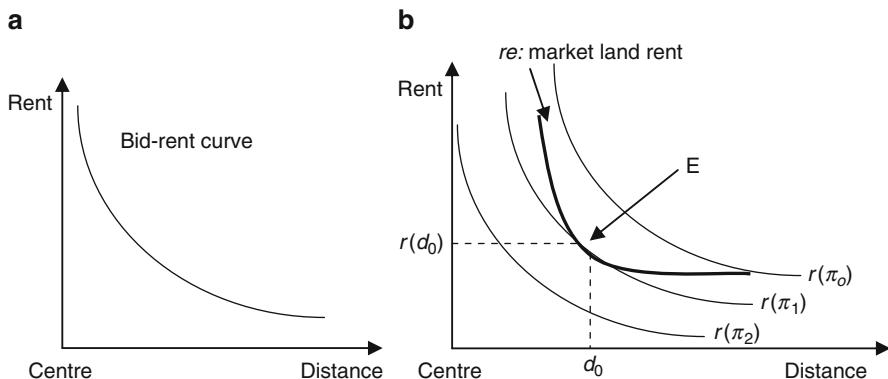


Fig. 27.2 The bid-rent curve and the location equilibrium for firms. (a) Bid-rent curve. (b) Location equilibrium for the firm

Equation (27.3) expresses the “bid rent” or the rent (by square meter) that the entrepreneur is willing to pay at differing distances from the center, once costs and the entrepreneur’s intended profit have been subtracted from revenues. Profits remaining equal, a more central location implies a willingness to pay higher rent because the entrepreneur incurs lower transportation costs and obtains higher revenues. Likewise, a suburban location can yield the same profit if and only if less rent is paid for the land: the saving on land cost must offset the higher transportation costs and the lower revenues that less central locations entail (Fig. 27.2a).

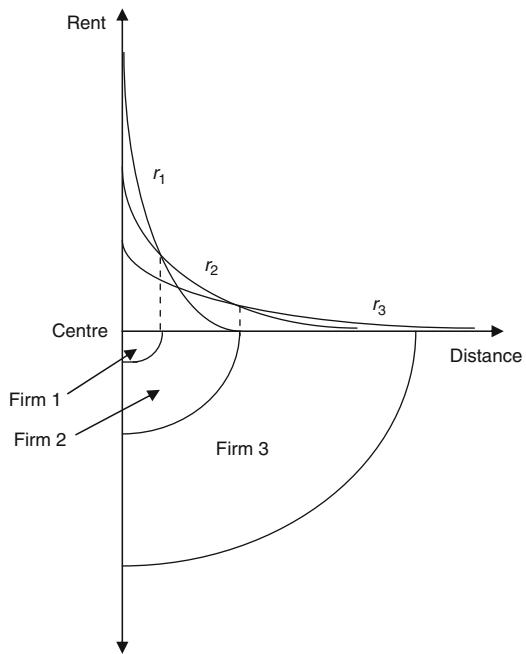
At every distance from the center (e.g., d_0 in Fig. 27.2b), if the firm wants to increase its profits, it must offer a lower rent. Vice versa, at the same distance, it can offer a higher rent if it is willing to accept lower profits. It is therefore possible to plot different bid-rent curves for an individual firm, all of them with the same slope and each of them defined on the basis of a different profit level which increases toward the origin (Fig. 27.2b).

In a partial equilibrium framework, which assumes as known the “market land rent curve” (i.e., the real market cost of land expressed by curve re in Fig. 27.2b), it is possible to define the optimal location for the firm. Along the market land rent curve (re), the firm will choose the location yielding the highest profit, which is expressed by the tangency of the market rent curve with the lowest bid-rent curve. In Fig. 27.2b, the location equilibrium is reached at point E and thus at a distance d_0 from the center and with a rent equal to $r(d_0)$ (Fujita 1985).

27.2.4 The Location of Activities in a City: A General Spatial Equilibrium Approach

If we discard the assumption that a city already exists (a market land exogenously determined) and therefore if we move away from the identification of an urban location for a new firm entering a city, we abandon a partial spatial equilibrium

Fig. 27.3 Location equilibrium for activities with different propensities for central location. Firm 1 = high propensity for central location. Firm 2 = average propensity for central location. Firm 3 = low propensity for central location. r_1 = rent bid by firm 1, r_2 = rent bid by firm 2, r_3 = rent bid by firm 3,



model and go back to von Thünen's general spatial equilibrium framework, which entails an interesting interpretation of the allocation of the urban space between alternative production activities or between production and residential activities.

Suppose the existence of a point in space (a center) attracting firms because of the presence of a market for their goods. Firms compete for locating closest to the central location in a homogeneous space around the center. Let us assume that there exist firms with different propensity for central location. The slopes of the bid-rent curves will differ according to the different levels of propensity for central location; as the propensity increases, firms will be willing to pay more for houses in order to locate (one unit of distance) closer to the center (Fig. 27.3).

The three firms are distributed across the urban area as in von Thünen's model: each area will be occupied by the firms that make the highest rent bid. Market land rent will be the envelope of the bid-rent curves at each distance from the center so that the city can be depicted as a set of concentric rings each containing the firm willing to pay the highest rent for that distance (Fig. 27.3).

But what determines the propensity for a central location? To reply to this question, a reasoning on the slope of the bid-rent curve, which expresses the variation in the cost of land due to a one unit of variation in the distance from the center, is helpful. The slope is given by

$$\frac{\partial r(d)}{\partial d} = (p_x - \pi - c(d)) \frac{\partial x(d)}{\partial d} - \frac{\partial c(d)}{\partial d} x(d) \quad (27.4)$$

Table 27.1 Taxonomy of activities with high propensity for central location

$\frac{\partial c(d)}{\partial d}$	$\frac{\partial x(d)}{\partial d}$	$p_x - \pi - c(d)$	x	Activities
Low	High	Normal	Normal	Commercial activities, shopping centers but also specialized shops, or luxury good shops
Low	Normal	High	Normal	Advanced service functions (e.g. lawyers, specialized doctors), or of activities that require a prestigious location
Low	Normal	Normal	High	Travel agencies, insurance brokers
High	Normal	Normal	Normal	Activities dependent on population and on other central activities
High	Normal	High	High	Financial intermediaries

Source: Camagni 1992

This shows that, at one unit of distance further away from the center, the rent offered to maintain the same profit level π diminishes because of increased transportation costs and decreased revenues. Equation (27.4) contains the four elements that, on their own or in combination, theoretically explain higher propensity of activities to central location; an activity will be in fact interested to locate in the center if:

- The costs of moving one unit of good toward the center ($\frac{\partial c(d)}{\partial d}$) are high.
- The influence of distance on the demand of goods ($\frac{\partial x(d)}{\partial d}$) is high.
- Extra profits ($p_x - \pi - c(d)$) are high.
- The value of activities per unit of land (x) is high.

Table 27.1 presents some examples of activities characterized by the major values of the slope of the bid-rent curve:

- (a) Activities oriented toward a high demand density, like commercial activities, shopping centers but also specialized shops, or luxury good shops, all characterized by a strong influence of distance on the demand of goods.
- (b) Advanced service functions (e.g., lawyers, specialized doctors) or of activities that require a prestigious location that can obtain thanks to their oligopolistic position (banks, insurances, public and private managerial functions), whose costs of moving one unit of good toward the periphery and the influence of distance on the demand of goods sold by unit of land are low but the extra profits of a central location are very high; through a central location, these activities abandon a perfect competition market and differentiate the product quality through the use of a traditional urban input factor, like information.
- (c) Travel agencies and insurance brokers, all characterized by a very high value of their activity per unit of land.
- (d) Activities that depend on a central market, with a high transportation cost of the final output: all industrial and service activities that depend on population and central activities.
- (e) Financial activities mostly linked to the stock exchange, whose propensity of a central location is the result of high transportation costs, monopolistic profits, and a high value of activity per unit of land.

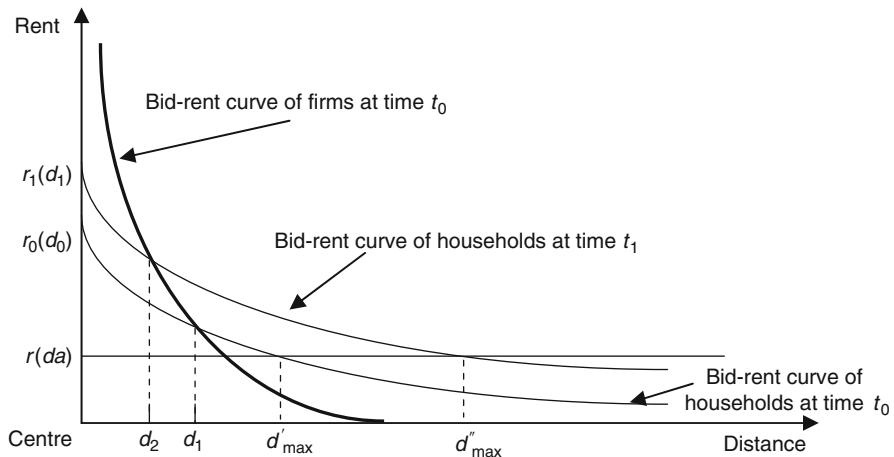


Fig. 27.4 Location equilibrium for households and firms. Legend: $r(d_a)$ agricultural land value

These reflections provide already a first evidence of how these models, strongly abstract in their nature, are able to describe conditions which closely match actual reality.

The same logic of Fig. 27.3 can be applied for a simultaneous analysis of the location of firms and households. On the hypothesis that the rent gradient of firms is higher than that of households (i.e., firms are willing to pay higher unit rents in order to move one unit of distance closer to the center, as it is usually the case in the real world), the bid-rent curves for firms and households will be those shown in Fig. 27.4.

This model leads to two important results. The first is that it identifies the bid-rent curves of firms and households endogenously. Let us assume that at time t_0 , households choose a level of rent $r_0(d_0)$ characterized by a certain level of utility. For equilibrium to come about, the level of utility must be such that it determines an amount of population and a labor supply equal to the labor demand of firms. If households too have chosen a high level of utility and therefore make rent bids which are too low, the population located in the city (in the range $d_1 - d'_\text{max}$) may be insufficient to satisfy the labor demand by firms. The availability of work will attract new households into the city, with a consequent increase in demand for urban land which pushes the bid-rent curve up to $r_1(d_1)$ in Fig. 27.4. The city will expand (d''_max) until labor-market equilibrium has been reestablished at a lower level of utility.

The second important result of this model is that it divides the urban area between productive and residential activities. Urban land will be allocated to the activities able to pay a higher rent for each distance from the center – as in von Thünen's model. In this case, the central areas will be occupied by firms, while households will be pushed toward suburban areas: a theoretical result which closely reflects what actually happens in reality.

As a final remark, although these models are highly abstract, owing to their unrealistic hypotheses (isotropic space, a city with a single center), they are able to

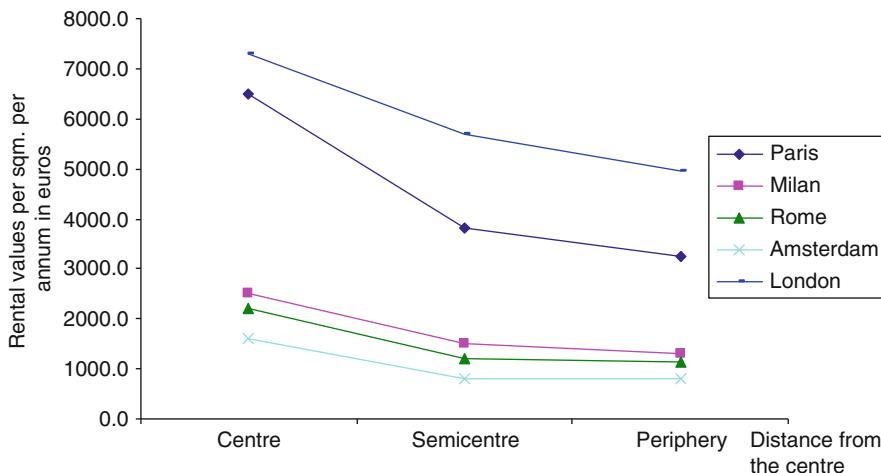


Fig. 27.5 Urban rent gradients in some major European cities (Rental values per sqm. per annum in euros – 2003) – Retail (Source: Capello 2007, p. 60)

describe conditions which closely match actual reality: an urban land rent gradient negative with respect to (business and management) broad suburban spaces for residential activities (Fig. 27.5).

27.2.5 Critical Remarks

Despite their logic, elegance, and economic rigor, these models have a number of theoretical elements which weaken their overall logical structure. One of them is the decisive role played by commuting in determining location equilibrium. If real behavior does not comply with the perfect rationality envisaged by the models so that commuting is of less importance for a person's utility, the entire theoretical-conceptual edifice collapses. However, this shortcoming can be partly remedied, however, if we acknowledge that the costs of transport to the center and the desire to reduce them may reflect other important aspects of an individual's utility function when she/he makes location choices, like accessibility to information, recreational services, and opportunities for social interaction.

A second shortcoming is more serious. These models concern themselves neither with how a city center is organized nor with what happens outside the city itself. They restrict themselves to interpreting locational behavior within the area extending between a hypothetical aspatial center and the physical boundary of the city. Moreover, when these models are used to interpret location equilibrium, not internally to a city but among cities, and therefore on the hypothesis that the city is part of an urban system and that firms may decide to relocate to other cities with attractively higher levels of utility or profit, they display a clear interpretative weakness. On the hypothesis that firms have equal production functions, there

can only be indifference to alternative locations in other cities if all these exhibit – in the logic applied here to describe them – the same bid-rent curve and the same boundary-rent curve and are therefore all of the same size.

If this is the case, there will be an urban system made up of cities which are all of equal size, but this circumstance is amply contradicted in the real world. In order to deal with this defect, the conceptual framework should be able to accommodate the hypothesis that locational advantages differ according to the size of the city and that rents – the monetary counterpart of the advantages that firms obtain from central urban locations – vary (distance from the center remaining equal) from city to city.

Only thus is it possible to conceive a location equilibrium with cities of different sizes. Yet this also requires acceptance of the idea that large, medium-sized, and small cities are structurally different and perform different functions in the overall economy and consequently have specific production specializations: a hypothesis at odds with the basic features of these models and which instead opens the way for the general equilibrium models discussed in ► [Chap. 3, “Labor Market Theory and Models”](#) of this handbook.

27.3 The Location of Industrial Activities in Space: Weber's Model

27.3.1 Transportation Costs and Agglomeration Economies

The aim of other kinds of location models is to explain location choices of firms by considering the two great economic forces that organize activities in space: transportation costs and agglomeration economies. These forces push the location process in opposite directions since they simultaneously induce both the dispersion and the spatial concentration of production.

It is because of agglomeration economies that spatial concentration comes about. Widely used in regional economics, the term “agglomeration economies” denotes all economic advantages accruing to firms from concentrated location close to other firms and result from the concentration of economic activities in space. However, there are two forces which work in the reverse direction and give rise to dispersed location. The first is the formation in the agglomeration area of increasing costs or diseconomies, these being (i) the prices of less mobile and scarcer factors (land and labor) and (ii) the congestion costs (noise and air pollution, crime, social malaise) distinctive of large agglomerations. These diseconomies are generated above a certain critical threshold. However, the second factor – transportation costs – is of greater interest because these costs countervail the spatial concentration of activities whatever level of agglomeration has been reached. For in conditions of perfect competition, perfectly mobile production factors, fixed raw materials, and demand perfectly distributed across the territory, the existence of transportation costs may erode the advantages of agglomeration until activities are geographically dispersed and the market becomes divided among firms, each of which caters to a local market.

The way in which these two opposite forces can be taken at the same time into account is elegantly presented in Weber's model, described hereafter. As we will see, unlike the location models described in Sect. 27.2, which identify just one factor organizing activity in space (land rent), Weber's model envisages a different location equilibrium according to the spatial principle that patterns activities in space (agglomeration economies rather than minimum transportation costs).

27.3.2 Weber's Model

One of the first and best-known studies on the spatial concentration of industry dates back to 1909. In that year, the economist Alfred Weber constructed an elegant location model where the costs of transportation among production site, raw materials markets, and the final goods market (which together define a minimum transportation cost) are directly compared against localization economies (Weber 1929). The prevalence of one element over another determines the geography of industry location.

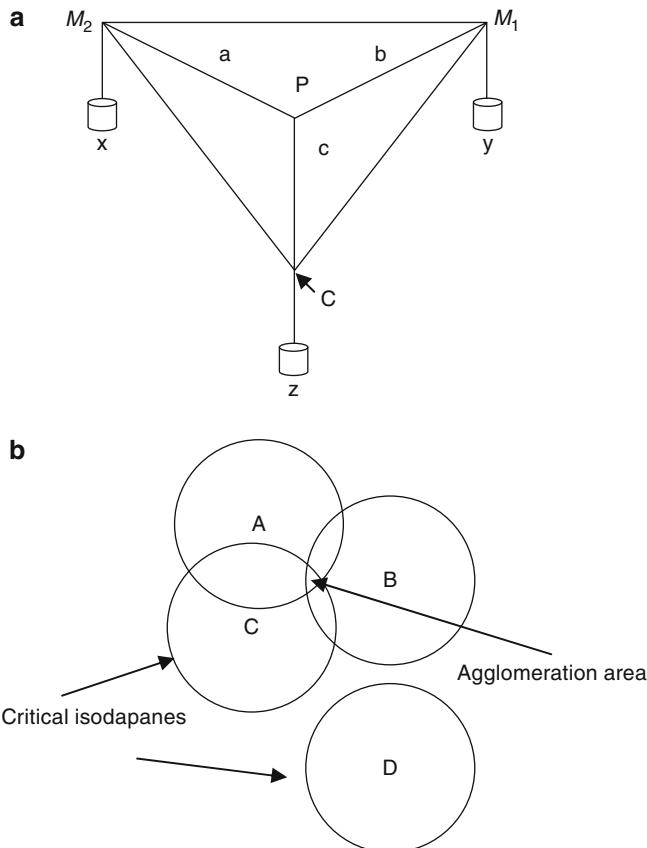
Weber's model is based on the following simplifying assumptions:

- (a) There is a punctiform market for the good (C in Fig. 27.6a).
- (b) Two raw materials markets, these too punctiform, are located at a certain distance from each other (M_1 and M_2 in Fig. 27.6a).
- (c) There is perfect competition in the market, i.e., firms are unable to gain monopolistic advantages from their choice of location.
- (d) Demand for the final good is price inelastic. Demand is said to be inelastic when the price of a good changes but the quantity of the good demanded (or supplied) varies less than proportionally or remains the same.
- (e) The same production technique is used in every possible location; production costs are therefore given and constant.

The location choice results from a complex calculation performed in two stages. In the first, the firm looks for the location that assures the minimum transportation cost between the production site, the raw materials market, and the final market for the good produced. In the second stage, the firm compares the advantages of agglomeration (localization economies) against the higher transportation costs that it would incur by choosing the new location instead of the one with minimum transportation costs.

The first stage of calculation identifies the location that assures minimum transportation costs. Let x and y be the tonnes of raw materials present, respectively, in markets M_1 and M_2 and required to produce one unit of output, and let z be the tonnes of the finished good to be transported to the final market C . Total transportation costs (CT) are expressed as a function of the weight of the good to be transported and the distance to cover:

$$CT = xa + yb + zc \quad (27.5)$$



Isodapane: geometric locus of points of constant additional cost of transportation with respect to the least-cost location

Fig. 27.6 Weber's location equilibrium. (a) The locational triangle: choosing the location with the minimum transportation costs. (b) The agglomeration areas. Legend: x and y tonnes of raw materials, z tonnes of final goods, M_1 , market of one raw material, M_2 market of another raw material, C market of the final good, A, B, C and D; different firms

where a , b , and c are, respectively, the distances in kilometers between the raw materials markets and the production site and between the latter and the final market and xa , yb , and zc represent the “forces of attraction” that push the firm, respectively, toward points M_1 , M_2 , and C (Fig. 27.6a).

The minimum cost location solution can be identified:

- At a point inside the triangle formed by joining M_1 , M_2 , and C if none of the “forces of attraction” exceeds the sum of the other two. In economic terms, this situation occurs when the cost of transporting the z tonnes of the good 1 km further away from the outlet market is less than the costs of transporting the x and y tonnes of raw materials 1 km further away from their source market.

- At corner C of the triangle, i.e., the final market, if the sum of the costs of transporting the x and y tonnes of raw materials 1 km further away from their market is less than the cost of transporting the z tonnes of final good produced one extra kilometer. This situation comes about because of the greater relative weight, in the composition of the finished product, of ubiquitous raw materials with respect to those that must be transported. Weber calls this condition “market oriented.”
- At a point closer to the raw materials markets if the sum of the costs of transporting the x and y tonnes of raw materials 1 km more is greater than the extra cost of transporting the z tonnes of the finished good. This situation can be explained by the lesser relative weight, in the composition of the final good, of ubiquitous raw materials with respect to localized raw materials and/or the product’s loss of weight during the manufacturing process. Weber calls this location “raw-material oriented.”

Weber provides a practical solution to the problem of identifying the minimum point. He hypothesizes a triangular board (the location triangle) in which three holes are drilled at the vertexes M_1 , M_2 , and C . Threads are passed through these holes (Fig. 27.6a), and their ends are knotted together on the upper surface of the board. Weights respectively proportional to x , y , and z are attached to the other ends of the threads below the board. The point at which the knot of the three threads lies on the upper surface of the board corresponds to the point of minimum transportation costs.

In the second stage of the location choice process, the firm compares the least-cost location with an alternative one where it can enjoy localization economies – for instance, the availability of labor at lower cost and/or better quality.

Assuming that P in Fig. 27.6a is the location point with the lowest transportation costs, Weber describes the “isodapanes”: curves along which the additional transportation cost that the firm must pay in order to cover a certain distance from the least-cost location remains constant. On the assumption that other firms operate in the same sector and that these firms obtain advantages from concentrated location such that they have a pecuniary advantage equal to v , the decision to relocate will be taken if and only if each firm’s isodapane measuring an extra transportation cost equal to the agglomerative advantage (v) intersects with the isodapanes of the other firms. In this case, in fact, within the area of intersection, the additional transportation costs are less than the advantages generated by concentrated location. In Fig. 27.6b, firms A , B , and C find themselves in this situation, and they relocate, but not so firm D , for which the agglomerative advantage is no greater than the additional transportation cost.

27.3.3 Criticisms of the Model

Weber’s model has made a permanent and major contribution to industrial location theory. Its principal merit is that it uses entirely rational modes of reasoning: for instance, comparison between the advantages of an alternative location and the

additional transportation costs that it would generate. Nevertheless, the model has a number of shortcomings:

- Its static nature. The model identifies the least-cost location on the basis of productive efficiency, but it ignores dynamic aspects such as innovation at the microeconomic level, while, at the macroeconomic one, it neglects changes in income distribution and in the relationships among agglomeration advantages, rents, and wages.
- Its transport-oriented nature. The cost of transportation defines first and foremost the most efficient location; only subsequently does it identify alternative locations. Some critics have claimed that this approach is less efficient than one based on the direct search for a point of minimum total production cost.
- Its abstractness, which makes the least-cost location difficult to calculate in real settings. It is rather unlikely, in fact, that the weight of raw materials in the final weight of the good can be calculated, distinguishing *inter alia* the weight of the raw materials to be transported from those present at the production site.
- Its nature as a partial equilibrium model which entirely neglects possible interactions among firms.
- Its supply-side bias. The criticism most frequently made of the model is that it is excessively oriented to the supply side: it makes no mention of demand factors, assuming that demand is unlimited and inelastic to price variations.

27.3.4 Strength of the Model

The main strength of the model is that for the first time it opened the way to reflections on the importance of agglomeration economies in location choices. The importance and power of agglomeration economies in the real world is evident. The number of concentrations of economic activity in space is very high. Local districts or clusters of SMEs denote a local area with a strong concentration of small- and medium-sized firms, each specialized in one or a few phases of the production process (or activities subsidiary to it) serving the needs of the area's principal sector. In this sense, they reflect Weber's concept of localization economies, i.e., those advantages stemming from the physical proximities of firms belonging to the same sector. In the real world, firms do not only cluster to achieve increased static efficiency of their production processes (i.e., an increase in firms' revenues or a decrease in their costs); firms get also *dynamic efficiency* in the form of increases in their innovative and creative capacity.

Cases in point are Silicon Valley in California, "Route 128" in the Boston area, Baden-Württemberg in the South of Germany, Jutland in Denmark, Småland in Sweden, and Sophia Antipolis close to Nice, to cite only some examples.

Weber's model does not explain the existence of agglomeration economies and gives them for granted. However, it has to be recognized that the intuition to highlight them as an important element in the choice of firms' location has opened the way to a long stream of theoretical reflections to interpret the formation of production agglomeration in space.

27.4 Conclusions

This chapter has surveyed the two groups of location theories developed to explain the determinants of location choices by industrial firms. First, it described models of a strictly neoclassical nature, which seek to account for the allocation of land between alternative activities within a spatial structure of uniform supply in space and a punctiform source of demand. High demand for access to central areas triggers competition among firms or between firms and households to obtain locations closer to the market, or more generally to a hypothetical central business district.

Land rent is the main factor that organizes activities in urban space. According to strict economic logic, competition for land closer to the center is resolved by its allocation to activities able to pay higher rents.

The virtues of these models are their rigor and their stringent economic logic. Their main weakness emerges when they set out to explain the location choices made by firms between cities with different levels of utility or profit. Indifference to alternative locations, which is the long-period equilibrium condition, is guaranteed if and only if cities offer the same utility and the same profit and therefore, according to the model's logic, if and only if cities are of the same size. Yet this implies the existence of an urban system consisting of cities which are all of the same size – a circumstance widely belied by reality. In order to understand the economic reasons for the existence of urban systems with cities of different sizes, consideration must be made of the functional characteristics of cities. This is an aspect which the models described thus far are unable to handle and which is instead addressed by the models discussed in the next chapter.

The second kind of models presented in this chapter is Weber's model. On the hypothesis that demand and supply structures are punctiform in space, the model elegantly and convincingly explains the existence of territorial agglomerations on the basis of two great economic forces which induce either the concentration or the dispersion of activities in space: agglomeration economies on the one hand and transportation costs on the other. Still today, these forces are components of more modern, and in certain respects more complex, models which seek to conjugate location choices with local growth dynamics (see ► Sect. 5, “Location and Interaction” of this handbook), and it is on the balancing of them that the geographical organization itself of activities depends.

References

- Alonso W (1960) A theory of the urban land market. *Pap Proc Reg Sci Assoc* 6:149–157
Alonso W (1964a) Location theory. In: Friedmann J, Alonso W (eds) *Regional development and planning: a reader*. MIT Press, Cambridge, MA, pp 78–106
Alonso W (1964b) Location and land use: towards a general theory of land rent. Harvard University Press, Cambridge, MA
Beckmann MJ (1969) On the distribution of urban rent and residential density. *J Econ Theory* 1(1):60–68

- Cairncross F (1997) *The death of distance*. Harvard Business School Press, Cambridge
- Camagni R (1992) *Economia Urbana: Principi e Modelli Teorici*. La Nuova Italia, Rome
- Capello R (2007) *Regional economics*. Routledge, London
- Fujita M (1985) *Urban economic theory: land use and city size*. Cambridge University Press, Cambridge, MA
- Isard W (1956) *Location and space-economy*. MIT Press, Cambridge, MA
- Muth R (1968) Urban residential land and housing market. In: Perloff H, Wingo L (eds) *Issues in urban economics*. The Johns Hopkins Press, London, pp 285–333
- Muth R (1969) *Cities and housing*. The University of Chicago Press, Chicago
- Ricardo D (1971) *Principles of political taxonomy and taxation* (orig edn 1817). Penguin Books, Hardmondsworth
- von Thünen JH (1826) *Der Isolierte Staat in Beziehung auf Landwirtschaft und Nationalökonomie*. Puthes, Hamburg
- Weber A (1929) *Alfred Weber's theory of the location of industries* (orig edn 1909). University of Chicago Press, Chicago. *Über der Standort der Industrien*. Verlag Mohr, Tübingen
- Wingo L (1961) *Transportation and urban land. Resources for the Future*, Washington, DC

Schools of Thought on Economic Geography, Institutions, and Development

28

Philip McCann

Contents

28.1 Introduction	528
28.2 New Economic Geography	528
28.3 The New Urban Agenda	532
28.4 The Evolutionary and Institutional School	533
28.5 Conclusions	536
References	537

Abstract

This chapter reviews some of major thematic approaches which have characterized urban and regional research over recent decades. Three broad schools of research are discussed, namely, the new economic geography, the new urban agenda, and the evolutionary and institutional school. The major assumptions underlying each of the schools of thought are outlined, and the broad areas of agreement and disagreement between the three schools of thought are highlighted. The changing economic realities on the ground in many regions, whereby the previously dominant large cities are no longer the key drivers of economic growth, pose major conceptual, analytical, and empirical challenges to all three of these schools of thought, schools which had emerged precisely during the period when major cities were reemerging as the drivers of growth.

P. McCann

Department of Economic Geography, University of Groningen, Groningen, The Netherlands
e-mail: p.mccann@rug.nl

28.1 Introduction

Recent developments in urban and regional economics have provided different discourses regarding the nature of regional and urban growth. These different discourses reflect different schools of thought, and these different schools of thought themselves can be interpreted in terms of what Lakatos (1970) defines as different “research programs,” programs which are built on self-contained sets of assumptions and their associated methodologies, programs which are largely internally consistent within their own assumptions, and programs which compete in terms of explanatory power with other research programs (McCann 2007). Over the last two decades, three major schools of thought can be identified in urban and regional economics, all of which arose initially in the late 1980s and early 1990s and which still continue to heavily influence urban and regional economic debates in rather different ways. Obviously, it is possible to categorize different research themes and schools diversely, but for the purposes of this chapter, the most parsimonious grouping is into three major research programs which can be listed as:

- (i) The new economic geography
- (ii) The new urban agenda
- (iii) The evolutionary and institutional school

These three research programs are in no way suggestive of an exhaustive listing of all regional science activities over the last two decades, nor are they entirely mutually exclusive of one another. However, they do, at least, capture many of the major lines of enquiry over recent years and also reflect many of the fundamental elements of agreement and disagreement between scholars regarding the workings of the interregional economic system and the lines for future required research. What these three research programs have done, however, is to reignite the debates regarding the role of cities and regions in national and international economic development and to reposition questions of geography back in the center of many wider debates about economic growth.

28.2 New Economic Geography

New economic geography, or NEG for short as it is known, arose initially out of the work of 2008 Nobel Laureate Paul Krugman, who applied the analytical approaches of new internal trade theory to questions of geography. In 1991, there appeared what is now widely regarded as being landmark publication (Krugman 1991) and which subsequently generated a wave of new analytical and empirical approaches to describing the geography of economics. Krugman’s early work with Masahisa Fujita and Anthony Venables (Fujita et al. 1999) has subsequently been extended and developed by such a wide range of scholars, to the extent that this research program has become a whole subfield of urban and regional economics in its own right.

As with all Lakatosian research programs, new economic geography is based on a particular set of assumptions. Although the models employed in this field can very quickly become rather complex and technical, in essence the models are based on very simple analytical framework, relating to the effects of variety, scale, factor mobility, and transactions costs. The first assumption concerns the relationship between size and variety, in that new economic models work on the assumption that the diversity of choices available to individual households and firms increases the welfare of those individuals and firms. As such the welfare effects of diversity are understood to operate primarily in two dimensions: firstly, the welfare of individuals and households is assumed to increase according to the range of consumers' choices and the consumption opportunities which are available to them; secondly, the profitability of a firm is assumed to be related to the variety of intermediate inputs among which a firm can choose to purchase supplies. In simple terms variety and choice increases (consumption or production) welfare, because it improves the quality of the matches between what the household or firm ideally wishes to consume and the opportunities to realize those choice preferences. In terms of geography, in new economic geography the primary setting where the elements of variety and size are naturally assumed to come together is in the context of a city or an urban area. The justification for this is based on the standard arguments relating to agglomeration economies, in that greater city size implies more variety of choices and better matching outcomes, for all individuals and firms in the city, as well as greater price competition and more efficient pricing outcomes.

The second element within new economic geography models is the question of factor migration. In early new economic geography models, labor is apportioned between employment in agricultural activities, which exhibit constant returns to scale, and manufacturing goods, exhibiting increasing returns to scale. The ability of factors (labor and capital) to move between sectors and places differs in different model specifications, based on the assumptions we make regarding the mobility or otherwise of factors. More recent models in new economic geography have also incorporated labor-commuting behavior and even knowledge and technological movements between places and sectors in their specifications.

The third and final element in the basic new economic geography framework is the specific way in which the distance costs are treated. New economic geography models specify distance costs or transport costs ways which are very different to traditional regional science approaches to these matters. Traditional region science models tend to specify product prices incorporating explicitly or implicitly transport costs in terms of functions which closely relate to f.o.b. (free on board) or c.i.f. (cost, insurance, freight) pricing schemes evident in transport economics, including ones that allow for frequency-optimization arrangements. In contrast, new economic geography models generally define transport costs or distance costs in a form known as an "iceberg" specification, whereby the value of the delivered good (or person in commuting models) falls with distance. The outcome of this iceberg specification is that the costs of distance are seen to increase at the margin with increasing distance. The reason for using this iceberg specification is to facilitate

the modeling of transport issues alongside the modeling of the other elements of the overall schema. This specification is loosely related to the von Thünen transport cost arguments, although the actual mechanics of, and justification for, adopting this approach are much less clear-cut and rather more problematic than is often assumed (McCann 2005).

In the new economic geography framework, a city is the natural setting where scale and variety come together. The production and consumption efficiencies associated with both scale and matching coincide in cities, and it is assumed that larger cities contain a greater variety of inputs and consumption choices, which provide the efficiency-scale effects. Taken together, these advantages therefore provide the “centripetal” forces which encourage factors to collocate together in cities, and the rate and scale of the colocation of factors depends in part on the mobility of the factors. In contrast, the iceberg transport costs, in which the costs associated with distance increase at the margin, provide the centrifugal forces favoring dispersion of factors. More recent new economic geography models also allow for localized congestion effects to provide another centrifugal force. While the centripetal forces encourage colocation and the formation of large cities which dominate their hinterlands, centrifugal forces allow more distance locations to maintain their own market hinterlands, rather than being dominated by the major urban centers. The overall observed spatial distribution of economic activities is argued to depend on the balance between these two opposing forces, with some regions in which economic activity is dispersed and other regions in which it is spatially concentrated.

There are numerous insights generated by new economic geography models, and more recently there are also new insights arising from the models which integrate new economic geography with new growth theory frameworks. However, for the purposes of this chapter, rather than trying to provide an exhaustive listing of all insights, it is useful to highlight a couple of key insights.

One of the most important new economic geography insights is provided by the model of Krugman and Venables (1995), in which two economies, one of which is large and contains a large agglomeration and one of which is small and contains only small urban areas, are increasingly integrated with one another due to falling trade costs, from a situation of having previously been largely closed to one another due to trade costs having been too high. In traditional neoclassical trade theory, such a process of mutual integration would have been assumed to provide standard factor price equalization-type outcomes, whereby factor movements and specialization and trade distributions would have continued until factor rewards are equalized in both countries. However, the Krugman and Venables (1995) model demonstrates that completely different results are possible. In particular, the benefits of the mutual trade openness are seen to generally favor the large economy.

The Krugman-Venables (1995) result, which is initially counterintuitive to many observers, is important in that it demonstrates the role played by geography in determining factor rewards. Specifically: when the two economies are largely closed to each other, they both exhibit similar production patterns, because the

high trade barriers encourage local production and a tendency toward autarky. However, as trade costs begin to fall, the greater agglomeration and home-market advantages of the large economy begin to tell, leading to an increasing center-periphery divergence in favor of the large economy. Moreover, this core-periphery pattern will be bolstered the greater is the mobility of the factors. In other words, falling trade costs will foster a greater concentration of activities in the larger economy, and these scale-productivity advantages of the large economy over the small economy will only disappear as trade costs fall toward zero, in which case colocation provides no advantages. Indeed, much of the evidence of falling trade barriers and falling communications costs over the last two decades suggests that exactly this type of process has taken place in many parts of the world, favoring core regions and large urban centers at the expense of other regions (McCann 2007).

A second key insight of new economic geography models is that the evolving spatial distribution of activities within a one-dimensional new economic geography framework (Fujita et al. 1999) can be seen to be broadly consistent with the types of hierarchical structures evident in both the urban-system patterns of Christaller and Lösch and also the rank-size rule pattern of urban size distributions. Moreover, the one-dimensional simulations of Fujita et al. (1999) have more recently been extended to two-dimensional simulations by Steldor (2005), and these clearly also demonstrate that clustering and dispersion are natural outcomes of spatial competition processes.

To say that the results of new economic geography models are largely consistent with the urban patterns and distributions generated by other rather more partial frameworks is not in any way to unequivocally demonstrate validity and veracity of the new economic geography models. Rather, what it does demonstrate is that the new economic geography models are well able to capture various different dimensions of the underlying mechanics of the interregional economic system and to do so in a manner which is reflective of a self-organizing economy operating in a general equilibrium type system. Indeed, the ability to capture these different dimensions simultaneously within the same schema represents both a major theoretical breakthrough on the part of new economic geography and also provides a major step forward in terms of how we frame questions about regional and urban issues, which nowadays increasingly start from the king of general equilibrium departure point of new economic geography models.

Having said this, it is also probably fair to say that the theoretical developments in new economic geography have been rather more significant than the empirical developments. Certainly, there are increasing numbers of spatial econometric models incorporating new economic geography frameworks into their analysis (Fingleton and Fischer 2010), but the primary impact of new economic geography has primarily been in terms of analytical insights. On this point, probably, the major empirical development in new economic geography relates to the ways in which the market potential can be estimated, and in particular to move beyond the standard market potential approach to one in which the impacts of competition, variety and wages are also included in market potential measurement (Redding and Venables 2004).

The weakness with these approaches is that they are seen to be very sensitive to the actual specifications employed (Bosker and Garretsen 2010), in which small changes in parameters and assumptions lead to major empirical changes.

28.3 The New Urban Agenda

The second major research program we discuss is what we will broadly define here as the *new urban agenda*, and this stream of research focuses on the investigation of the different mechanisms driving urban agglomeration economies (Glaeser 2011). In general, a major difference between this approach and that of new economic geography is that while new economic geography is overwhelmingly a theoretical research program with much more limited empirical elements embedded within it, the new urban agenda is fundamentally an empirical research program with some theoretical elements embedded within it.

The paper which is generally accredited with galvanizing the new urban agenda research program was by Glaeser et al. (1992), in which the role played by industrial specialization and diversity in fostering urban economic growth was examined empirically. The arguments in this chapter tend to center on the role played by economies of localization which are spillovers evident in the same sectors in the same places (or “MAR-externalities,” named after Marshall-Arrow-Romer) or economies of urbanization (or Jacobs externalities) which are spillovers evident between sectors in the same places. This seminal Glaeser et al. (1992) paper was followed by numerous other contributions since. More recent extensions of this approach also include the role played by cities as centers of consumption as well as production (Glaeser et al. 2001; Shapiro 2006), in which urban amenities are also considered as part of the attractor forces encouraging agglomeration. The overall conclusion from the US papers is that across all regions, sectoral diversity and urban scale appear to be better for fostering the long run growth of a region than sectoral specialization (Glaeser and Gottlieb 2009), and this conclusion clearly chimes with the basic assumptions of new economic geography. On the other hand, however, internationally there appears to be no such consensus on the growth-enhancing role of sectoral specialization or diversity, with some research pointing to the advantages of specialization and others to the advantages of diversification (de Groot et al. 2009). As such, it appears to depend somewhat on the context, whereby the degree of a region’s diversity or specialization may play a different role in different locations and different time periods.

Given the applied econometric focus of these papers, this research program has also highlighted the importance of having both good urban data and good data analysis. Various indices have been constructed in order to try to capture the extent of spatial clustering both within sectors and between sectors, as an indirect way of capturing different types of agglomeration effects. Such indices need to allow for the fact that some observations of clustering are likely to be entirely purely due to statistical reasons which are unrelated to actual firm location behavior or choices but rather to structural or cartographical issues. For example, several indices have

been developed to control for the fact that some observed patterns of employment clustering and dispersion may simply be the result of the distribution of firm sizes (Guimarães et al. 2007). Other indices have also been developed to control for the effects of the drawing of cartographical boundaries (Duranton and Overman 2005) and the effects of shared linkages between industries (Ellison et al. 2010). In each of these cases, the aim of these indices is to account for the degree of spatial clustering which occurs in addition from what would be expected on the basis of the definitions of firms, areas, or industries.

The new urban agenda research program has spurred the detailed econometric and theoretical work on the micro-foundations of urban areas (Duranton and Puga 2001). The role of cities as engines of economic growth has been emphasized, and, along with new economic geography, this research program has greatly helped to position geographical issues at the center of wider discussions of economic growth and particularly in the case of the USA (Glaeser and Gottlieb 2009).

One likely future change of emphasis, however, is being signaled by the recent work of the OECD (2009a, b, 2011), which shows that the growth-enhancing role of major cities in many countries has largely faded and that the growth contribution of such cities, while being important, is also nowadays largely unchanging. Increasingly, in many countries, national economic growth is more associated with smaller and medium-size centers than with the largest centers. At the same time, the concept of a “representative” region is actually declining, with regional heterogeneity increasing. Moreover, if smaller cities tend to be more specialized than larger cities, then this recent evidence would also appear to point in some cases to some reemerging advantages of urban sectoral specialization over diversification. These empirical findings also appear to be robust to different empirical specifications and are also broadly comparable across countries. Given that the new urban agenda research program emerged out of early-1990s empirical observations suggesting that urban scale and diversity were reemerging as a driver of economic growth, what these recent OECD empirical findings now imply in terms of shifting discourses regarding the role of cities remains to be seen.

28.4 The Evolutionary and Institutional School

A rather different research program to the two already discussed is that of the evolutionary and institutional school. This approach takes as its departure point the fact that places are not just by geographical distance, proximity, and accessibility but also by “knowledge,” distance, and accessibility (Boschma 2005) and that understanding the means by which knowledge is transmitted and mediated is critical. The argument here is that regions and sectors are related to each other according to various different dimensions which may not be best captured by the types of models and empirics adopted by either the new economic geography or the new urban agenda research programs. In particular, this research program stresses that many regions are “related” to each other in technological or institutional terms as well as geographical terms, and this has important

implications in terms of our understanding of how innovation and growth processes operate between regions.

Both the new economic geography and the new urban agenda frameworks can be linked to the earlier endogenous growth work, whereas the evolutionary and institutional school focus more on the insights of authors such as Aghion and Howitt (1992), who argue that economic growth is primarily driven neither by efficient factor allocation and pricing nor by competition, specialization, or variety. Slightly differently, the Aghion and Howitt (1992) framework is based on the assumption that growth is driven fundamentally by Schumpeterian processes of creative destruction, characterized by risk taking, entrepreneurship, and innovation. These Schumpeterian-type models do allow for the knowledge spillover, input variety, and human capital elements which are evident in other endogenous growth models and also in the new economic geography and new urban agenda frameworks. However, in these Schumpeterian models, competition is assumed to take place in an environment where technological change allows for major and fundamental shifts in the nature of competition. These shifts in competition, which are driven by innovation, imply that economic growth may follow quite different trajectories, depending on the technological breakthroughs which take place and the innovations associated with these breakthroughs. As such, there is not necessarily any preordained growth rate or growth trajectory to which the economy is assumed to converge, because it depends on how well positioned an economy is to take advantage of the newly emerging technologies. This line of research therefore emphasizes issues such as the diffusion of knowledge, the problems of technological “lock-in,” and the institutional capability of the economy, and in many ways these approaches mirror and build on the arguments put forward by Porter (1990) and Saxenian (1994).

When translated into the context of geography and regions, this line of reasoning implies that differences in regional growth are largely attributed to differences in the technological profile of different places, and these in turn depend on the interactions between history and geography. Such interactions are assumed to be the result of evolutionary processes, in which technologies and institutions play the equivalent role of genes in biology. These different technological profiles are assumed to lead to differences in the ability of regions to connect with, and to relate to, the newly emerging and most competitive technologies, and there are two aspects of this. Firstly, the ability of region to link with and exploit the technological developments in other leading regions is also argued to depend on the degree of technological congruence of the regions. Secondly, the ability of a region to take advantage of newly emerging technologies within its own productive capacity is assumed to depend on the region’s prevailing industrial structure.

On the first point, the assumptions of the evolutionary approach imply that the ability of one region to “learn” from another region depends on their degree of technological congruence between the regions. “Proximity” in terms of technological profiles facilitates knowledge spillovers and exchanges and thereby overcomes many of the problems associated with geographical distance. On the other hand, distance in terms of technology limits spillovers and knowledge exchanges, irrespective of the geographical distance.

A second, and arguably even more important insight of these evolutionary approaches, is that a region's ability to grow is conditioned by its own technological development trajectory – in other words, by its own technological history. More specifically, the argument here is that a region's growth potential is likely to be stronger if it aims to diversify into technological fields which are closer and more related to the region's current dominant technologies than in other largely unrelated technological fields. This is because such closely related technological shifts are assumed to better allow regional assets to exploit and build upon the region's existing skills and capabilities. This "related variety" (Frenken et al. 2007) argument, which is a combination of Schumpeterian and Darwinian thinking, therefore implies that the most promising trajectories or pathways forward for a region's growth are found by diversifying into technologies which are closely related to the existing dominant technologies, and there is now a large and growing body of empirical evidence to support this argument (Boschma and Iammarino 2009). Moreover, the evidence of this takes various forms in that both the inflows of new firms and also the founding of new local firms are found to be higher in related technological fields, while the outflows of firms or firm failures are also less likely in these same closely related sectors. Moreover, these effects prove to be even stronger at the regional scale than at the national scale. These Schumpeterian-Darwinian types of arguments all suggest that the clues to differing regional performance are related to questions of how knowledge flows are mediated and argues that technological profiles both reflect and also mediate the region's ability to acquire, generate, and process knowledge.

Other institutional approaches also adopt a similar line of reasoning regarding historical trajectories but from a slightly different point of view. Since the early 1990s, a great deal of new thinking about development and growth has emphasized the role played by institutions and governance systems in economic development. For many governance systems, an overreliance on a highly centralized state precludes the widespread engagement by local stakeholders (Barca et al. 2012), the result of which is only very limited endogenously driven local development. In contrast, an overreliance on decentralized systems often leads to coordination failures, rent seeking, duplication, and an absence of any coordinated strategy. Beyond this, however, it has become widely accepted that ability of an institutional system to facilitate growth and development is dependent not only on a "good" architectural design of the system but fundamentally also on the ways in which all of the system's actors, stakeholders, and interested parties interact with each other. In the case of institutions, one of the most important developments in our understanding centers on the role of what is known as social capital (Putnam 1993), whereby social capital relates to all of the social norms, social rules, and social conventions that operate within a society. Although the concept is almost a century old (Westlund 2009), it was reinvigorated by two seminal books written by Putnam (1993, 1996) which brought the arguments into contemporary mainstream economics and political science, and the arguments are particularly pertinent for urban and regional systems (Westlund 2009).

Putnam's (1993) initial work was based on the major development differences within Italy, which he ascribed to longstanding historical differences in the regional levels of trust in governance institutions. Putnam's argument is that these

historically determined levels of social capital have very long-lasting implications. Broadly, the less that individuals trust the protection to be afforded them from the formal legal and government institutions, the lower will be the levels of economic development, because weaker formal institutions make entrepreneurial activities increasingly risky. In contrast, high degrees of trust in well-functioning institutions mean that more people are willing to undertake entrepreneurial risks. As such, social trust and economic activity are argued to be highly correlated over time, and the long-term development of a region is therefore assumed to be closely related to the institutional history of the region.

As well as heavily determining a region's trajectory via the quality and efficacy of its institutional arrangements, however, social capital also plays a role in the ongoing development of region by influencing the ability of the region to adapt to changes (Putnam 1996). On this point, social capital can be understood in terms of two distinct types, namely, *bonding* capital and *bridging* capital, whereby bonding capital relates to the connection between people of similar types while bridging social capital refers to the ability of people to relate to other types of people. The former appears to be important in terms of keeping communities locally cohesive and may be an important element in ensuring that localities are "resilient" in times of adversity, whereas the latter appears to be important in helping localities to adjust to new economic realities.

In terms of our earlier arguments, it may well be the case that large and diverse urban areas display higher levels of bridging social capital, and this may facilitate greater knowledge flows and learning opportunities for adjusting and diversifying. Indeed, this may be one of the major economic advantages of diversity, a point also picked up on by Florida (2002). In contrast, smaller and more geographically peripheral centers may exhibit greater bonding social capital, which while being positive in terms of maintaining resilience in times of economic hardship, may limit the growth potential in stronger periods. The institutional and social history of the locality therefore becomes crucial for understanding the economic evolution of the region, and if these arguments are also linked to the related variety arguments, together the evolutionary and institutional school point to the technological, institutional, and social profile of the region as crucial for understanding its growth patterns, performance, and potential, over and above questions of variety, diversity, specialization, or scale, all of which are central to both the new economic geography and new urban agenda research programs.

28.5 Conclusions

The differences in approach between these different research programs or schools of thought do imply some differences in terms of how we think about the regional economy. On the one hand, the related variety, institutional, and social capital arguments tend to emphasize specific systemic factors within the local economic system as determining regional growth, and these themes closely mirror the themes picked up on by the work on new industrial areas, industrial districts, and clusters.

In contrast, the new economic geography and the new urban agenda approaches tend to see largely system-wide features as playing the dominant role. However, all three of these schools of thought, and in particular the new economic geography and new urban agenda schools, now face the challenge of responding to the recent empirical observations of the OECD (2009a, b, 2011) and MGI (2011), which suggest that the types of locations which these approaches emphasize, namely, the largest core cities, in reality may be rather less important for future economic growth than their analytical centrality implies. Moreover, the fact that both successful regions and struggling are becoming increasingly heterogeneous even within their own groupings (OECD 2009a, b) means that there is no simple typology describing an optimum structure. In reality, the actual outcomes in any particular region are likely to be a mixture of both region-specific and also the system-wide effects (Shearmur and Polèse 2007; Doloreux and Shearmur 2012). What is not yet clear is exactly how the balance between region-specific effects and system-wide effect, or between local influences and global influences, will change over the coming decades. The current literature was primarily forged during a period in which large cities were increasingly reaping the rewards of globalization, but in many countries this pattern nowadays appears to be changing (Barca et al. 2012), and this itself may warrant some changes in the lines of major enquiry.

References

- Aghion P, Howitt P (1992) A model of growth through creative destruction. *Econometrica* 60:323–351
- Barca F, McCann P, Rodriguez-Pose A (2012) The case for regional development intervention: place-based versus place-neutral approaches. *J Reg Sci* 52(1):134–152
- Boschma RA (2005) Proximity and innovation: a critical assessment. *Reg Stud* 39(1):61–74
- Boschma RA, Iammarino S (2009) Related variety, trade linkages and regional growth. *Econ Geograph* 85(3):289–311
- Bosker M, Garretsen JH (2010) Trade costs in empirical new economic geography. *Pap Reg Sci* 89(3):485–511
- de Groot HLF, Poot J, Smit M (2009) Agglomeration externalities, innovation and regional growth: theoretical perspectives and meta-analysis. In: Cappello R, Nijkamp P (eds) *Handbook of regional growth and development theories*. Edward Elgar, Cheltenham, pp 256–281
- Doloreux D, Shearmur R (2012) Collaboration, information and the geography of innovation in knowledge intensive business services. *J Econ Geograph*, 12(1):79–105
- Duranton G, Overman HG (2005) Testing for localization using micro-geographic data. *Rev Econ Stud* 72(4):1077–1106
- Duranton G, Puga D (2001) Nursery cities: urban diversity, process innovation, and the life cycle of products. *Am Econ Rev* 91(5):1454–1477
- Ellison G, Glaeser EL, Kerr WR (2010) What causes industry agglomeration? Evidence from coagglomeration patterns. *Am Econ Rev* 100(3):1195–1213
- Fingleton B, Fischer MM (2010) Neoclassical theory versus new economic geography: competing explanations of cross-regional variation in economic development. *Ann Reg Sci* 44(3):467–491
- Florida R (2002) *The rise of the creative class*. Basic Books, New York
- Frenken K, Van Oort FG, Verburg T (2007) Related variety, unrelated variety and regional economic growth. *Reg Stud* 41(5):685–697

- Fujita M, Krugman P, Venables AJ (1999) The spatial economy. MIT Press, Cambridge
- Glaeser EL (2011) Triumph of the city: how our greatest invention makes us richer, smarter, greener, healthier, and happier. The Penguin Press, New York
- Glaeser EL, Gottlieb JD (2009) The wealth of cities: agglomeration economies and spatial equilibrium in the United States. *J Econ Lit* 47(4):983–1028
- Glaeser EL, Kallal HD, Scheinkman JA, Shleifer A (1992) Growth in cities. *J Polit Econ* 100(6):1126–1152
- Glaeser EL, Kolk J, Saiz A (2001) Consumer city. *J Econ Geograph* 1(1):27–50
- Guimarães P, Figueiredo O, Woodward D (2007) Measuring the localization of economic activity: a parametric approach. *J Reg Sci* 47(4):753–774
- Krugman P (1991) Increasing returns and economic geography. *J Polit Econ* 99(3):483–499
- Krugman P, Venables AJ (1995) Globalization and the inequality of nations. *Q J Econ* 110(4):857–880
- Lakatos I (1970) Falsification and the methodology of scientific research programmes. In: Lakatos I, Musgrave A (eds) Criticism and the growth of knowledge. Cambridge University Press, Cambridge, pp 91–196
- McCann P (2005) Transport costs and new economic geography. *J Econ Geograph* 5(3):305–318
- McCann P (2007) Observational equivalence? Regional studies and regional science. *Reg Stud* 41(9):1209–1222
- MGI (2011) Urban world: mapping the economic power of cities. McKinsey Global Institute. http://www.mckinsey.com/insights/mgi/research/urbanization/urban_world
- OECD (2009a) How regions grow: trends and analysis. Organisation for Economic Cooperation and Development, Paris
- OECD (2009b) Regions matter: economic recovery, innovation and sustainable growth. Organisation for Economic Growth and Development, Paris
- OECD (2011) Regional outlook 2011. Organisation for Economic Cooperation and Development, Paris
- Porter ME (1990) The competitive advantage of nations. Free Press, New York
- Putnam R (1993) Making democracy work: civic traditions in modern Italy. Princeton University Press, Princeton
- Putnam R (1996) Bowling alone: the collapse and revival of American community. Simon & Schuster, New York
- Redding SJ, Venables AJ (2004) Economic geography and international inequality. *J Int Econ* 62(1):53–82
- Saxenian A (1994) Regional advantage: culture and competition in silicon valley and route 128. Harvard University Press, Cambridge
- Shapiro JM (2006) Smart cities: quality of life, productivity, and the growth effects of human capital, *Rev Econ Stat*, 88(2):324–335
- Shearmur R, Polèse M (2007) Do local factors explain local employment growth? Evidence from Canada, 1971–2001. *Reg Stud* 41(4):453–471
- Stelder D (2005) Where do cities form? A geographical agglomeration model for Europe. *J Reg Sci* 45(4):657–679
- Westlund H (2009) Regions and the knowledge economy. Springer, Berlin/Heidelberg/New York

New Economic Geography: Past and Future **29**

Carl Gaigné and Jacques-François Thisse

Contents

29.1	Introduction	540
29.2	Cities and Manufacturing Firms	544
29.2.1	Agglomeration and Commuting Costs	545
29.2.2	The Decentralization of Jobs Within Cities	552
29.3	Cities and Services	554
29.3.1	Cities as Local Service Providers	555
29.3.2	The Size and Industrial Structure of Cities	556
29.4	The Future of Cities	561
29.4.1	Cities in Aging Nations	561
29.4.2	Are Compact Cities Ecologically Desirable?	564
29.5	Conclusions	566
	References	567

Abstract

This chapter does not aim to survey what has been accomplished in new economic geography (NEG) since the publication of Paul Krugman's seminal paper. Rather, we provide an overview of recent developments in the NEG literature that build on the idea that the difference in the economic performance of regions is explained by the behavior and interactions between households and

C. Gaigné
INRA, UMR1302 SMART, Rennes, France
e-mail: carl.gaigne@rennes.inra.fr

J.-F. Thisse (✉)
CORE, Université catholique de Louvain, Louvain-la-Neuve, Belgium
CMSE, NRU-Higher School of Economics, Saint-Petersburg, Russia
e-mail: jacques.thisse@uclouvain.be

firms located within them. This means that we consider NEG models which take into account land markets, thereby the internal structure and industrial mix of urban agglomerations.

29.1 Introduction

Ever since the publication of Krugman's (1991) pioneering paper, new economic geography (NEG) has given new life to spatial economics, which since then has made enormous progress by any previous yardstick. The very name “new economic geography” seems chosen to stir a debate: Is NEG economic geography proper or rather spatial economics? And is there anything really new in it? To the best of our knowledge, no economist before Krugman had been able to show how regional imbalances can arise within the realm of general equilibrium theory. To achieve this, Krugman has borrowed concepts and tools developed in modern economic theory, especially the Dixit and Stiglitz (1977) model of monopolistic competition, which is the workhorse of new growth and trade theories. As for transport costs, Krugman uses the iceberg technology: Only a fraction of a good shipped between two places reaches the destination, the missing share having melted on the way. This ingenious modeling trick, due to Samuelson (1954), allows integrating positive shipping costs without having to deal explicitly with a transport sector. Hence, Dixit, Stiglitz, and Samuelson form the trinity under which Krugman has combined increasing returns, commodity trade, and the mobility of production factors within his now famous “core-periphery” model.

In NEG, the distribution of activities emerges as the unintentional outcome of a myriad of decisions made by firms and households pursuing their own interest. Thus, *methodologically, NEG belongs to mainstream economics*. This is probably what distinguishes most NEG from economic geography proper. Our choice to focus on NEG only does not reflect any prejudice on our part. It is mainly driven by the need to stress how this approach can be used to highlight old and new issues. Being deeply rooted in mainstream economics, NEG has strong connections with several branches of modern economic theory, including industrial organization and international trade, but also with the new theories of growth and development. This permits cross-fertilizations which have been out of reach for a long time. We also want to stress that differences between alternative approaches are often overemphasized. Indeed, NEG and evolutionary economic geography share many common results (Jovanovic 2009). Furthermore, in terms of its subject matter, NEG cannot be considered alien to regional science and geography. Moreover, many ideas and concepts NEG builds on have been around for a long time, both in economics and regional science (Ottaviano and Thisse 2005). For example, the fundamental idea that the interplay between different types of scale economies and transport costs is critical for the way the space economy is organized, and so at various spatial scales (cities, regions, countries, continents), was known (at least) since the work of Weber and Lösch. It is fair to say, however, that those ideas were fairly disparate and in search of a synthesis.

It is now widely recognized that Krugman's main contribution was his entirely different and new approach to the origin of regional imbalance. Under constant returns, firms find it profitable to disperse their production to bring it closer to customers, as this will reduce transport costs without lowering productive efficiency. Such a space economy is the quintessence of self-sufficiency: If the distribution of factor endowments is uniform, the economy reduces to a Robinson Crusoe-type economy where each person produces for his or her own consumption. Under these circumstances, only differences in endowments of immobile production factors can explain the marked differences in the spatial distribution of activities and hence the need for interregional and international trade.

To a large extent, relying on first nature to explain the existence of large urban agglomerations and sizable trade flows amounts to playing Hamlet without the Prince. Krugman squarely tackles this problem by assuming that firms operate under increasing returns and imperfect competition on the product market. Such a combination is orthogonal to the standard paradigm of constant returns and perfect competition, which has dominated mainstream economic theory for a long time. Furthermore, to the trade-off between increasing returns and transport costs, Krugman (1980) has added a third variable: the size of spatially separated markets. The main accomplishment of NEG has been to highlight *how market size interacts with scale economies internal to firms and transport costs to shape the space economy*.

In NEG, the market outcome stems from the interplay between a dispersion force and an agglomeration force operating within a full-fledged general equilibrium model. In Krugman (1991) and Fujita et al. (1999), the dispersion force stems from the spatial immobility of farmers whose demands for the manufactured good are to be met. The agglomeration force is more involved and requires a more detailed description. If a larger number of manufactures are located in one region, the number of varieties locally produced is larger too. Then, manufactured goods are available at lower prices in the larger region because local varieties are cheaper than imported varieties. This in turn induces consumers living in the smaller region to move toward the larger region, where they may enjoy a higher standard of living. The resulting increase in the numbers of consumers creates a larger demand for the manufactured good, which, therefore, leads additional firms to locate therein. This implies the availability of more varieties in the region in question but less in the other because there are scale economies at the firm's level. Consequently, as noticed by Krugman (1991, p.486), *circular causation à la Myrdal* (1957) is present because these two effects reinforce each other: "manufactures production will tend to concentrate where there is a large market, but the market will be large where manufactures production is concentrated." The great accomplishment of Krugman was to integrate all these various effects within a unified framework and to show that the *level* of transport costs is the key-determining factor for the organization of the space economy.

When transport costs are sufficiently low, Krugman (1991) showed that manufactures are concentrated in a single region that becomes the *core* of the economy, whereas the other region, called the *periphery*, supplies only the agricultural good.

Firms are able to exploit scale economies by selling more in the larger market without losing much business in the smaller market. For exactly the opposite reasons, the economy displays a symmetric regional pattern of production when transport costs are high: Because the local markets are now protected by geographical separation, firms relax competition by being dispersed across regions. Hence, the core-periphery model allows for the possibility of convergence or divergence between regions, whereas the neoclassical model, based on constant returns and perfect competition, would predict convergence only. It is worth stressing that the dual structure made of a core and a periphery is brought about by market forces only as it is obtained in a setting formed by two regions that are *ex ante* identical. These results hold true in more general settings such as those discussed in great detail in Fujita et al. (1999) and Baldwin et al. (2003).

By focusing on the interactions between the product and labor markets, Krugman's work remains in the tradition of international trade. Although we recognize the limits of this approach, we believe it delivers a powerful framework which has proven to be very useful in dealing with a large number of issues. Yet, in this chapter, we do not discuss the canonical NEG models and their vast number of extensions. Rather, we provide an overview of recent developments in the NEG literature, which fits better the space economy of developed economies. In particular, we build on the idea that the difference in the economic performance of regions is, to some extent, explained by the behavior and interactions between households and firms that are located within them. *The focus thus shifts from the nation-state to the city region.* Therefore, we discuss NEG-like models in which the internal structure and industrial mix of urban agglomerations are determined in interaction with a land market.

To be precise, we start by focusing on the causes and consequences of the internal structure of cities because the way they are organized has a major impact of the well-being of people. In particular, housing and commuting costs, which we call *urban costs*, account for a large share of consumers' expenditures. For example, in the United States, housing accounts on average for 20 % of household budgets while 18 % of total expenditures is spent on car purchases, gasoline, and other related expenses which do not include the cost of time spent in traveling. In 2000, the total cost of people's journeys inside the Paris metropolitan area amounted to a staggering 34.3 billion euros, which is just over 8 % of the local GDP. As for housing, the price per square meter is, on average, 80 % higher in Paris than in the rest of France. This leads us to concur with Helpman (1998) for whom urban costs are the main dispersion force at work in modern urbanized economies.

In this alternative setting, an agglomeration is structured as a monocentric city in which firms gather in a central business district. Competition for land among consumers gives rise to land rent and commuting costs that both increase with population size. In other words, our approach endows regions with an urban structure which is absent in standard NEG models. As a result, the space economy is the outcome of the interaction between two types of mobility costs: the transport costs of commodities and the commuting costs borne by workers. The results presented in Sect. 29.2 for a monocentric city differ from those obtained by

Krugman: *the evolution of commuting costs within cities, instead of transport costs between cities, becomes the key-factor explaining how the space-economy is organized.* Moreover, despite the many advantages provided by the inner city through an easy access to highly specialized services, the abyssal fall in communication costs has led firms or developers to form enterprise zones or edge cities (Henderson and Mitra 1996). We then go one step further by allowing firms to form secondary business centers. This analysis shows how polycentricity allows reducing urban costs, which in turn permits a large city to retain its dominant position by accommodating a large share of activities.

Another change of focus is on services rather than manufactures. The bulk of the NEG literature has concentrated on manufacturing sectors, although employment in modern cities is found mainly in firms providing nontradable consumption (b2c) services. While the Industrial Revolution fostered the emergence of manufacturing cities, services continue to show a taste for cities that manufacturing sectors no longer have (Bairoch 1985). As stressed by Glaeser et al. (2001), the success of a city depends more than before on its role as a center of consumption, that is, on the supply of local amenities and services. Even though NEG conveys the image of an economy formed by an urban core hosting manufactures and a rural hinterland specialized in agriculture, a pattern now obsolete in developed countries, recent contributions to NEG pay more attention to the role played by local services in urban development. In Sect. 29.3, we recognize that services (e.g., health care, restaurants, and movie theaters) are conditioned on precise locations and study their intercity distribution. We then add a new dimension to the above analysis by blending a service sector and a manufacturing sector in an economy in which workers display both sectoral and spatial mobility.

These ideas are presented through the lenses of a new framework. NEG being to some extent a collection of specific examples, we have no reason to apologize for using another specific model, that is, the linear model of monopolistic competition proposed by Tabuchi and Thisse (2006), which is much easier to handle than CES-based models. As for the transport cost, it is added to the production cost and measured in terms of the numéraire. This avoids imposing binding relationships between prices and shipping costs. By yielding linear equilibrium conditions, this model delivers a full analytical solution that captures in a simple way the pro-competitive effects associated with market size and market integration. To accomplish this task, we use a NEG-like model that takes into account the following fundamental aspects of urban development: (i) cities can be monocentric or polycentric, (ii) cities supply nonhousing services as well as tradable goods, and (iii) cities have heterogeneous demographic structures involving different types of individuals (e.g., workers and retirees) who are attracted by different location factors. Moreover, the framework we use displays enough versatility to tackle new issues which are difficult to cope with by using the standard framework of NEG. We are well aware that the reader accustomed to NEG might be surprised by our choice of menu. It is worth stressing that the basic model used in this chapter can replicate the main results obtained by Krugman and others. It thus belongs to NEG. The seemingly different approach followed here has been chosen in the hope

of convincing the skeptical regional scientist that NEG is a lively field that still has a high potential for future research.

The remainder of this chapter reflects the above methodological choices. The basic model is presented in Sect. 29.2, and the market outcome is compared to Krugman's core-periphery structure. We show how commuting costs and population density impact on the location of economic activities *between* and *within* cities. The subsequent section focuses on nontradable consumption services and their interactions with tradable manufactured goods. To illustrate the potential of NEG in the study of policy-driven problems, Sect. 29.4 addresses the implications of an aging population for the urban system and the environmental and economic consequences of compact cities characterized by a high population density. Section 29.5 concludes.

29.2 Cities and Manufacturing Firms

The economy is formed by two regions, labeled $r = A, B$, and populated by $L > 0$ consumers who are free to choose where to live and work. To ease the burden of notation, we choose the unit of labor for $L = 1$. Unlike the standard core-periphery model where regions are pretty much spaceless places, we recognize explicitly that any sizeable human settlement takes the form of a city where economic agents compete for land; cities are assumed to be anchored and separated by a given physical distance.

There is one manufacturing sector and three goods: land (housing), a produced good, which is differentiated and tradable, and an unproduced homogeneous good, which is the numéraire. Space is one-dimensional and the opportunity cost of land is zero. Each region can be urbanized by accommodating firms and consumers according to rules described below. Whenever a city exists, it has a central business district (CBD) where firms set up. Because NEG has nothing really new to add to the reasons explaining why CBDs exist, it is convenient to assume that CBDs preexist.

As discussed in the introduction, the main reason for the existence of cities is the presence of increasing returns. Under scale economies internal to firms, consumers have direct access to the locally produced varieties, the number of which depends on the size of the local market. When they display a love for variety or when the city population is formed by individuals having each idiosyncratic tastes for ideal varieties, consumers are also inclined to consume varieties produced in other places. This in turn prompts trade in differentiated commodities across spatially separated urban markets. As observed by Hicks (1969, p. 56): “The extension of trade does not primarily imply more goods. . . . The variety of goods is increased, with all the widening of life that entails. There can be little doubt that the main advantage that will accrue to those with whom our merchants are trading is a gain of precisely this kind.” However, foreign varieties must be imported at a positive transport cost, which tends to make them more expensive.

The standard thought experiment of NEG is well known: How do firms and consumers locate when the cost of shipping the manufactured good *between* regions/cities steadily decreases? Once we account for a description of the space economy that fits better the contemporary world, this thought experiment must be supplemented by another one, that is, the impact of commuting costs *within* cities. In sum, both the transport costs of commodities and the commuting costs of people must be taken into account to understand how economic activities are distributed across space.

29.2.1 Agglomeration and Commuting Costs

We assume that the lot size is fixed and normalized to 1 where the tallness (i.e., the number of floors) of buildings is given by $\delta > 0$ regardless of their location in the city. As a consequence, the parameter δ is the *population density* which also measures the city's compactness. Consequently, when the total population of city r is λ_r , the city is described by an interval of length λ_r/δ . At the residential equilibrium, all consumers reach the same utility level. If land is available on both sides of the CBD, the residential equilibrium involves a symmetric distribution of consumers around the CBD with city r 's right-hand side limit at

$$\bar{x}_r = \frac{\lambda_r}{2\delta}$$

A. Consumers. Consumers share the same quasi-linear preferences, which imply that the land ownership structure has no impact on our results. Denoting by n the total mass of varieties, the utility derived from consuming q_i units of variety $i \in [0, n]$ is given by

$$u(q_i) = \alpha q_i - \frac{\beta}{2} q_i^2 - \frac{\gamma}{2n} q_i \int_0^n q_j dj \quad (29.1)$$

The parameters α , β , and γ are interpreted as follows: $\alpha > 0$ measures the desirability of the manufactured good with respect to the numéraire and $\gamma > 0$ is the degree of substitutability between variety i and any other variety, whence a higher γ means that varieties are less differentiated. The parameter β expresses the desirability of variety i with respect to the total consumption: A high value of β means that a consumer aims at equalizing her consumption over the entire range of varieties. This parameter therefore measures the intensity of consumers' love for variety. Moreover, Eq. (29.1) shows that the marginal utility of variety i decreases with its own consumption as well as with the total consumption of the manufactured good.

Preferences are obtained by nesting the subutility Eq. (29.1) into a linear utility:

$$U(q_0, q_{i \leq n}) = q_0 + \alpha \int_0^n q_i di - \frac{\beta}{2} \int_0^n q_i^2 di - \frac{\gamma}{2n} \int_0^n q_i \left(\int_0^n q_j dj \right) di \quad (29.2)$$

where q_0 is the quantity of the numéraire. The lot size being fixed, there is no need to specify how housing enters into preferences. In what follows, we ease the burden of notation by adopting two normalizations which entail no loss of generality: The unit of the numéraire is chosen for $\alpha = 1$ and the unit of the manufactured good for $\gamma = 1$ to hold.

Each consumer supplies inelastically one unit of labor. Consumers commute to the CBD where jobs are located and earn an income w_r which is determined at the equilibrium. The unit commuting cost is given by $t > 0$, and thus, a consumer located at $x > 0$ bears a commuting cost equal to tx units of the numéraire. In addition, each consumer is endowed with $\bar{q}_0 > 0$ units of the numéraire, which is sufficiently large for the individual consumption of the homogeneous good to be strictly positive at the equilibrium outcome. Hence, the budget constraint faced by a consumer living in city r is given by

$$\int_0^n q_{ir} p_{ir} di + q_0 + R_r(x)/\delta + tx = w_r + \bar{q}_0 \quad (29.3)$$

where p_{ir} (q_{ir}) is the price (consumption) of variety i in city r and w_r the wage paid by the firms set up in city r 's CBD. In this expression, $R_r(x)$ is the land rent at x and thus $R_r(x)/\delta$ the price paid by a consumer to reside at x .

A consumer chooses her location and consumption bundle so as to maximize her utility Eq. (29.2) subject to the budget constraint Eq. (29.3). This yields the following demand for variety i :

$$q_{ir} = \frac{1}{\beta + 1} - \frac{p_{ir}}{\beta} + \frac{1}{\beta(\beta + 1)} \bar{p}_r \quad (29.4)$$

where

$$\bar{p}_r = \frac{1}{n} \int_0^n p_{jr} dj$$

is the average price prevailing in city r . The demand Eq. (29.4) captures in the very simple way the impact of competition on a firm's demand: A higher (lower) average price shifts upward (downward) the demand for variety i because local competition is softer (tougher), thus making variety i more (less) attractive to city r -consumers.

In what follows, we call "urban costs" the sum of housing and commuting costs borne by a city r -consumer residing at any location x :

$$UC_r(x) = R_r(x)/\delta + tx$$

Let $\Psi_r(x)$ be the highest price a worker is willing to pay to reside at location x in city r . Because there is only one type of labor, the equilibrium land rent is such that

$$R_r^*(x) = \delta \max\{\Psi_r(x), 0\}$$

The lot size being fixed, a marginal hike in the commuting trip must be equal to the decrease in the bid rent Ψ_r , that is, $\partial\Psi_r/\partial x + t = 0$. Hence, $\Psi_r(x) = k - tx$ where k is a constant. Since the land rent at \bar{x}_r is equal to the opportunity cost of land, here zero, we have $k = t\bar{x}_r$, and thus, $\Psi_r(x) = t\lambda_r/2\delta - tx$. Therefore, the price paid to reside at x is the mirror image of the corresponding commuting costs:

$$R_r^*(x)/\delta = t\left(\frac{\lambda_r}{2\delta} - x\right) \quad \text{for } x < \bar{x}_r \quad (29.5)$$

Hence, the price paid by a consumer to live at $\bar{x} > 0$ decreases with the population density since the average commuting is shorter.

The urban costs borne by consumers in city r do not depend their residential location within this city and are equal to

$$UC_r = \frac{t\lambda_r}{2\delta} \quad (29.6)$$

Because they increase with city size, *urban costs act as the dispersion force*. As expected, intercity differences in urban costs increase with commuting costs and decrease with population density.

It remains to close the model by specifying the structure of land ownership. Unless explicitly mentioned, we assume for simplicity that the aggregate land rent is distributed to absentee landlords.

B. Producers. Firms produce a differentiated and tradable good under monopolistic competition and increasing returns; for simplicity, they do not use land. A firm produces a single variety and any two firms supply two differentiated varieties. Producing a variety of the manufactured good requires a fixed number ϕ of labor units. Hence, the total mass of varieties supplied in the economy is given by $n = 1/\phi$ and the mass of firms producing in city r by $n_r = \lambda_r/\phi$. So a lower value of ϕ means a higher labor productivity. Note that λ_r is also the share of firms located in city r .

The manufactured good is shipped between cities at the cost of $\tau > 0$ units of the numéraire. Markets are segmented, that is, each firm is able to set a price specific to the market in which its output is sold (Engel and Rogers 2001). Because preferences and technologies are symmetric, firms sell their varieties at the same price in each city. Thus, we may disregard the index i and write the profits earned by a city r -firm as follows ($s \neq r$):

$$\pi_r = p_{rr}q_r(p_{rr})\lambda_r + (p_{sr} - \tau)q_s(p_{rs})\lambda_s - \phi w_r$$

where p_{rr} is the price set by the local firms and p_{rs} the delivered price charged by the foreign firms; q_r and q_s are the quantities sold by the r -firms and s -firms, respectively.

The average price in city r is given by

$$\bar{p}_r = n_r p_{rr} + n_s p_{sr}$$

Plugging this expression into Eq. (29.4) and solving the first-order conditions yield the equilibrium prices:

$$p_{rr}^* = \frac{2\beta + \tau\lambda_s}{2(2\beta + 1)} \quad \text{and} \quad p_{rs}^* = p_{ss}^* + \frac{\tau}{2} \quad (29.7)$$

Both prices p_{rr}^* and p_{rs}^* capture the pro-competitive effects associated with a larger number of local competitors and lower transport costs. In other words, the prices of local and imported varieties are lower in large cities than in small ones. This result runs against the conventional wisdom which holds that tradables are more expensive in larger cities because land rents and wages are higher therein. Note, first, that this argument overlooks the fact that, given the continual flows of new goods, the consumer price index for urban consumers almost completely ignores the quality improvements of existing goods and the introduction of new goods which allows consumers to substitute low-priced goods for high-priced goods. Controlling for these effects, Handbury and Weinstein (2011) use a dataset covering 10–20 million purchases of grocery items and find that prices for the same goods are indeed significantly lower in larger cities. This highlights a trade-off which has been neglected in the urban economics literature: *consumers bear higher urban costs in larger cities but the tradable goods are supplied at lower prices*.

Furthermore, Eq. (29.7) shows that trade exacerbates competition in each city though the consumer (c.i.f.) price of imported varieties is higher than that of domestic varieties because distant firms have to cover the cost of shipping their output. Therefore, consumption is biased toward locally produced goods. By contrast, the producer (f.o.b.) price of imported varieties is smaller than that of local varieties. There is freight absorption to facilitate the penetration of varieties produced in distant places. Last, for intercity trade to occur and its pro-competitive effects to become concrete, transport costs cannot be too high: $p_{rs}^* - \tau > 0$. This condition holds regardless of the spatial distribution of firms if and only if $\tau < \tau_{trade} \equiv 2\beta/(2\beta + 1) < 1$.

Urban labor markets are local while labor market clearing implies that the creation and destruction of firms is governed by the location of consumers. Specifically, the equilibrium wage is determined by a bidding process in which potential firms compete for workers by offering them higher and higher wages until no firm can profitably enter the market. Put simply, operating profits are completely absorbed by the wage bill. The equilibrium quantities sold are given by $q_{rr}^* = p_{rr}^*/\beta$ and $q_{rs}^* = (p_{rs}^* - \tau)/\beta$. Plugging the equilibrium prices and quantities into π_r and solving for w_r give the equilibrium wage in city r :

$$w_r^* = \left[\lambda_r (p_{rr}^*)^2 + \lambda_s (p_{rs}^* - \tau)^2 \right] / \phi$$

Ottaviano and Thisse (2002) have shown that w_r^* increases (decreases) at a decreasing (increasing) rate with λ_r when ϕ is large (small) as well as when τ is small (large). In other words, the equilibrium wage rises with the size of the local

market when the labor productivity is high, shipping goods is cheap, or both. This implies that a higher wage need not be associated with a larger city. Such a result conflicts with the widespread idea that a higher employment density is associated with higher wages (Combes et al. 2008; Puga 2010): Standard estimates of the density elasticity of wages typically range from 0.02 to 0.05. However, one should keep in mind that these estimates are averages across cities. Moreover, if the existence of agglomeration economies is well documented, the literature has not succeeded yet to identify the relative importance of the channels through which they percolate.

Last, observe that the size of the product and labor markets is endogenous when consumers are mobile. Indeed, when consumers move from one city to the other, they bring with them both their production and consumption capacities. As a consequence, both the numbers of consumers and workers change.

C. The Formation of Manufacturing Clusters. The locational choice made by a consumer is driven by the indirect utility level she reaches in a city:

$$V_r(L_r) = CS_r + w_r^* - UC_r + q_0^* \quad (29.8)$$

where CS_r is the consumer surplus evaluated at the equilibrium prices and q_0^* is the equilibrium consumption of the numéraire. Hence, when choosing the city where she lives, a consumer takes into account the income she earns, the level of urban costs she bears, and the consumer surplus she enjoys in the city. Thus, though the individual demands Eq. (29.4) are unaffected by income, the migration decision takes income into account. Everything else equal, workers are pulled by the higher wage region. The population becoming larger, the local demand for the manufactured good is raised, which attracts additional firms.

Although the present framework differs from Krugman's (1991), it captures the same effects. It also encapsulates the following fundamental trade-off, which is absent in Krugman: Concentrating people and firms in a small number of large cities minimizes the cost of shipping commodities among urban areas but makes work trips (as well as many other within-city trips) longer; when dispersion prevails, consumers bear lower commuting costs but goods are more expensive because each city produces a small number of varieties and shipping them to the other cities is costly. Thus, both configurations give rise to specific spatial costs.

The economy is in equilibrium when no consumer has an incentive to change place. Denoting by λ the endogenous share of consumers residing in city A , a *spatial equilibrium* arises at $1/2 \leq \lambda^* < 1$ when the utility differential $\Delta V(\lambda^*) \equiv V_A(\lambda^*) - V_B(\lambda^*) = 0$. When $\Delta V(1) \geq 0$, $\lambda^* = 1$, and thus, all consumers and firms are set up in city A . Thus, location choices exhibit strategic complementarity (substitutability) when the $\Delta V(\lambda)$ is increasing (decreasing). NEG models typically display several spatial equilibria. In such a context, it is convenient to use stability as a selection device since an unstable equilibrium is unlikely to happen. An interior equilibrium is *stable* if, for any marginal deviation away from the equilibrium, the incentive system provided by the market brings the distribution of consumers back to the original one. This is so if and only if the slope

of the utility differential ΔV is strictly negative at λ^* . By contrast, an agglomerated equilibrium is stable whenever it exists.

Replacing each term of V_r by its expression leads to the following utility differential:

$$\Delta V(\lambda) = -\left[\frac{t}{\delta} - \frac{\Lambda(\tau)}{\phi}\right]\left(\lambda - \frac{1}{2}\right) \quad (29.9)$$

where

$$\Lambda(\tau) \equiv \frac{\tau[4\beta(3\beta+2) - (6\beta^2 + 6\beta + 1)\tau]}{2\beta(2\beta+1)^2}$$

with $\Lambda(\tau) > 0$ because τ is smaller than τ_{trade} .

It follows immediately from Eq. (29.9) that $\lambda = 1/2$ is always a spatial equilibrium. This equilibrium is stable when t exceeds $\delta\Lambda(\tau)/\phi$. Otherwise, the manufacturing sector is concentrated into a single city. As a result, *when commuting costs steadily decrease, there is a transition from dispersion to agglomeration*. The intuition behind this result is straightforward. When t is large, urban costs are sufficiently high to prevent the emergence of a big city. By contrast, there is agglomeration when t is small because the gains from variety overcome the land market crowding effect. Note also that increasing the population density δ amounts to decreasing the level of commuting costs. Hence, a high population density, a high labor productivity, or both makes agglomeration more likely. This is because a larger city allows individuals to consume a wider range of varieties priced at a lower level.

Finally, note that the catastrophic nature of the bifurcation obtained both here and in Krugman (1991) is an artifact due to the assumption of identical consumers. Once it is recognized that consumers are heterogeneous in their migration behavior, the transition becomes smooth (Tabuchi and Thisse 2002). Therefore, the interest generated by the result of sudden urbanization is unwarranted.

Though very simple, the above model allows understanding the role played by commuting costs in shaping the space economy. Consumers having a love for variety, they are attracted by the city supplying the wider range of local varieties, which are cheaper to buy than the imported varieties. By moving to this city, consumers increase the size of the local market, which makes local competition tougher. However, migration flows crowd out the land market and raise the urban costs borne by consumers residing in this city. Eventually, market clearing and labor mobility balance these various forces and select a spatial pattern involving either two small cities or one large city.

Note the difference with Krugman (1991): Here low transport costs are associated with the dispersion of activities. Indeed, when τ is very small, we have $\Lambda(\tau) \approx 0$, which implies $\delta\Lambda(\tau) - t\phi < 0$. Consequently, firms and consumers are located in two small cities. This is because consumers have more or less the same access to the whole range of varieties but obviate paying high urban costs through dispersion. This means that lowering transport costs induces the (partial)

de-industrialization of large manufacturing cities and the relocation of manufacturers in small cities or even in rural areas.

On the contrary, when τ is large and slightly smaller than τ_{trade} , $\Lambda(\tau)$ takes on its largest value so that $\delta\Lambda(\tau)/\phi$ is more likely to exceed t . Indeed, when transport costs are high, the agglomeration of the manufacturing sector allows consumers to have direct access to all varieties at a low price while firms are able to better exploit scale economies. In other words, high transport costs are likely to be associated with the agglomeration of activities.

To sum up, a drop in the cost of shipping commodities fosters the spatial decentralization of jobs and production: *Krugman's prediction is thus reversed*. This difference in results is simple to explain. In the above model, urban costs rise when consumers join the larger city, which strengthens the dispersion force. Simultaneously, lowering transport costs facilitates intercity trade. Combining the two forces tells us why dispersion arises. By contrast, in the core-periphery model developed by Krugman (1991), the spatial concentration of workers does not generate any cost in the core. Furthermore, the dispersion force stems from immobile farmers who live in what becomes the periphery. This force gets weaker when farmers can be supplied at a lower cost. Consequently, manufacturing firms choose to locate in the same region to benefit from a larger market. Krugman's conclusions thus hold in our setting provided that commuting costs are low and a sufficiently large share of consumers is immobile.

D. The Bell-Shaped Curve of Spatial Development. The above analysis suggests that the way the space economy is organized depends on the interplay between transport and commuting costs. Historically, it is well known that both costs have fallen at an unprecedented pace (Bairoch 1985). Therefore, what matters is the relative evolution of these two types of costs. For a long time, high transport costs have been the main impediment to trade. Even though the report of the "death of distance" is premature, it is clear that, within developed countries, the cost of shipping commodities has reached today a level which is much lower than commuting costs, which remain relatively high. As a consequence, the main dispersion force no longer lies in the cost of supplying distant markets, but in the level of urban costs. Under these circumstances, we may speculate that, though economic integration has initially fostered a more intensive agglomeration of economic activities, its continuation is liable to generate a redeployment of activities that could lead to a kind of geographical evening out. In short, one may expect the process of spatial development to unfold according to a *bell-shaped curve*.

To be precise, agglomeration occurs during the second phase of the integration process. The dispersion in the first and third integration phases emerges for very different reasons. In the former phase, the manufacturing sector is dispersed because shipping its output is expensive and, in the latter phase, because the smaller city has comparative advantage in terms of urban costs. Simply put, the relationship between economic integration and spatial inequality is not monotone: *while the first stages of economic integration exacerbate regional disparities, once a certain threshold is reached, additional integration starts undoing them* (for a more detailed discussion of the bell curve, see Combes et al. 2008).

29.2.2 The Decentralization of Jobs Within Cities

As seen above, globalization could well challenge the supremacy of large cities, the reason being that the escalation of urban costs would shift employment from large monocentric cities to small cities where these costs are lower. However, this argument relies on the assumption that cities have a monocentric morphology. The main point we wish to stress here is that decentralizing the production of goods in *secondary business districts* (SBD) may allow large cities to retain a high share of firms and jobs. Under these circumstances, firms are able to pay lower wages while retaining most of the benefits generated by large urban agglomerations. For example, Timothy and Wheaton (2001) report substantial variations in wages according to intra-urban location (15 % higher in central Boston than in outlying work zones, 18 % between central Minneapolis and the fringe counties). As they enjoy living on larger plots and/or move along with firms, consumers may also want to live in suburbia. Consequently, the creation of sub-centers within a city, that is, the formation of a *polycentric city*, appears to be a natural way to alleviate the burden of urban costs. It is, therefore, no surprise that Anas et al. (1998, p. 1442) observe that “polycentricity is an increasingly prominent feature of the landscape.”

For the redeployment of activities in a polycentric pattern to happen, firms set up in SBDs must be able to maintain a good access to the main urban center, which requires low communication costs. For example, about half of the business services consumed by US firms located in suburbia are supplied in city centers. By focusing on urban and communication costs, we recognize that both agglomeration and dispersion may take two quite separate forms because they are now compounded by the *centralization* or *decentralization* of activities within the same city. Such a distinction is crucial for understanding the interactions between cities and trade.

A. Polycentric Cities. We build on Cavaillès et al. (2007) and extend the above model by allowing manufacturing firms to locate in the CBD or to form a SBD on each side of the CBD. Both the CBD and the SBDs are surrounded by residential areas occupied by consumers. Because the higher-order services are still provided in the CBD, firms established in a SBD must incur a communication cost $K > 0$ so that the profit of a firm located in a SBD is given by $\pi_r - K$ whereas π_r is the profit of a firm established in CBD. In what follows, the superscript C is used to describe variables related to the CBD, whereas S describes the variables associated with a SBD.

Denote by y_r the right endpoint of the area formed by residents working in the CBD and by z_r the right endpoint of the residential area on the right-hand side of the SBD, which is also the outer limit of city r . Let x_r^S be the center of the SBD in city r . It is easy to show that these points are given by

$$y_r = \frac{\theta_r \lambda_r}{2\delta} \quad x_r^S = \frac{(1 + \theta_r)\lambda_r}{4\delta} \quad (29.10)$$

where $\theta_r < 1$ is the share of jobs located in the city r -CBD.

At a city equilibrium, each worker maximizes her utility subject to her budget constraint, each firm maximizes its profits, and markets clear. Individuals choose their workplace (CBD or SBD) and their residential location for given land rents and wages in the CBD (w_r^C) and in the SBD (w_r^S). The wage wedge between the CBD and a SBD is given by

$$w_r^C - w_r^S = t(2y_r - x_r^S) = \frac{t}{\delta} \frac{3\theta_r - 1}{4} \lambda_r \quad (29.11)$$

where we have used the expressions for y_r and x_r^S given in Eq. (29.10). In other words, the difference in the wages paid in the CBD and in the SBD compensates exactly the worker for the difference in the corresponding commuting costs. Moreover, the wage wedge is positive as long as $\theta_r > 1/3$, that is, the size of the CBD exceeds the size of each SBD. Note also that a larger population in city r raises the wage wedge. Indeed, as the average commuting cost rises, firms located in the CBD must pay a higher wage to their workers.

Within each workplace (CBD or SBD), the equilibrium wages are determined by a bidding process in which firms compete for workers by offering them higher wages until no firm can profitably enter the market. Hence, the equilibrium wages are related through the following expressions: $w_r^{C*} = w_r^*$ and $w_r^{S*} = w_r^* - K/\phi$. Given these equilibrium wages and the location of workers, firms choose to locate either in the CBD or in a SBD. At the city equilibrium, no firm has an incentive to change place within the city and no worker wants to change her working place and/or her residence.

Substituting w_r^{C*} and w_r^{S*} into Eq. (29.11) and solving with respect to θ_r yields

$$\theta_r^* = \min \left\{ \frac{1}{3} + \frac{4\delta K}{3t\phi\lambda_r}, 1 \right\} \quad (29.12)$$

which exceeds $1/3$ as long as $K > 0$. Clearly, the city is polycentric ($\theta_r^* < 1$) if and only if

$$K < \frac{t\phi\lambda_r}{2\delta}$$

The higher the communication costs, the lower the commuting cost, or both, the larger the CBD. In the limit, both SBDs shrink smoothly and the city becomes monocentric. In contrast, a larger population fosters the emergence of a polycentric city.

B. The Emergence of Polycentric Cities. The utility differential between cities now depends on the degree of job decentralization within each city. The indirect utility of an individual working in the CBD is still given by Eq. (29.8), but the urban costs Eq. (29.6) are replaced by the following expression:

$$UC_r \equiv \frac{t\lambda_r}{2\delta} \theta_r^*$$

Everything else equal, urban costs take on lower values when jobs are decentralized into the SBDs. As a consequence, the existence of SBDs allows the large cities to maintain their primacy.

The utility differential Eq. (29.9) becomes

$$\Delta V(\lambda) = -\left[\frac{t}{3\delta} - \frac{\Lambda(\tau)}{\phi}\right]\left(\lambda - \frac{1}{2}\right)$$

when both cities are polycentric and

$$\Delta V(\lambda) \equiv -2\left[\frac{2t}{3\delta} - \frac{\Lambda(\tau)}{\phi}\right]\lambda + \left[\frac{t}{\delta} - \frac{\Lambda(\tau)}{\phi} - \frac{4K}{3}\right]$$

when only one city is polycentric ($\theta_1^* < \theta_2^* = 1$).

Unlike standard models but as in Cavailhès et al. (2007), the economy displays a richer set of stable equilibrium configurations: (i) dispersion with two identical monocentric cities, (ii) agglomeration within a single monocentric city, (iii) partial agglomeration with one large polycentric city and a small monocentric city, (iv) agglomeration within a single polycentric city, and (v) dispersion with two identical polycentric cities. Once communication costs are low enough, the economy traces out the following path when the ratio t/δ steadily decreases. By inducing high urban costs, a high t/δ -ratio leads to the dispersion and decentralization of jobs, that is, the economy involves two polycentric cities. When δ gets higher or t lower, urban costs decrease sufficiently for the centralization of jobs within one city to emerge at the market outcome. However, urban costs remain high enough for the equilibrium to involve two cities having different sizes and structures. Last, when the t/δ -ratio takes on very low values, urban costs become almost negligible, which allow saving the cost of shipping the manufactured good through the existence of a single city.

The multiplicity of stable equilibria has also an important implication that has been overlooked in the literature: *different types of spatial patterns may coexist under identical technological and economic conditions*. It should be no surprise, therefore, to observe different types of urban systems in the real world.

29.3 Cities and Services

In Sect. 29.2, as in most NEG models, consumers have access to the entire range of produced varieties. As observed by Handbury and Weinstein (2011), residents of larger cities have, ceteris paribus, access to more varieties than residents of smaller cities. The rising share of nontradable consumption services explains, to some extent, this fact. What distinguishes service cities from the manufacturing cities is that the cost of shipping local services are prohibitive. Consequently, consumers have access only to the varieties produced in the city in which they live.

29.3.1 Cities as Local Service Providers

To start with, we consider a setting with no manufacturing sector and focus on the impact of commuting costs on the spatial distribution of nonhousing services. The circumstances in which one large city or two small cities emerge are the issue discussed in this section.

Consumer preferences are given by Eq. (29.2), except that the set of available varieties in city r is now given by n_r instead of n . The profits earned by a city r -firm are given by

$$\pi_r = p_r q_r(p_r) \lambda_r - w_r \phi$$

Because service firms compete only on their local market, the equilibrium price of a city r -variety is obtained by setting $\lambda_s = 0$ in Eq. (29.7):

$$p_r^* = \frac{\beta}{2\beta - 1} \equiv \mathbf{p} \quad (29.13)$$

which is the same in the two cities. Observe that a stronger love for variety yields a higher market price because service firms have more market power.

The consumer surplus generated by a single variety is equal to $S = (1 - \mathbf{p})^2 / \beta$, which is independent of the city size. Because the value of S does not play any role in the analysis undertaken here, we set $S = 1$. As for the total surplus, it is equal to the number $n_r = \lambda_r / \phi$ of locally produced varieties, which increases with both the city size and the labor productivity. Put simply, *consumers living in larger cities have access to more nontradable services*.

The urban labor markets being local, the equilibrium wage paid by firms established in city r is equal to

$$w_r^* = \mathbf{p}^2 \lambda_r / \phi$$

In other words, wages are higher in larger cities because the local market is bigger. Observe that this correlation does not reflect a difference in well-being. As expected, w_r^* also increases when workers are more productive because more firms compete on urban labor markets.

Replacing each term of V_r by its expression leads to the following utility differential:

$$\Delta V(\lambda) = - \left[\frac{t}{\delta} - \frac{2(1 + \mathbf{p}^2)}{\phi} \right] \left(\lambda - \frac{1}{2} \right)$$

As in Sect. 29.2.1, the symmetric pattern ($\lambda^* = 1/2$) is always a spatial equilibrium. However, when $t/\delta < 2(1 + \mathbf{p}^2)/\phi$, this equilibrium is unstable because the utility differential is positive for all values of λ . The market outcome therefore involves a single large city accommodating all consumers ($\lambda^* = 1$). Thus, *even in the absence of trade, consumers and firms may choose to be agglomerated within a single large city*. This is so when (i) commuting costs are low, (ii) the population density is high, and (iii) the array of local services is wide. The intuition is fairly

straightforward. By being agglomerated in a single city, consumers have access to all varieties. Furthermore, low fixed costs favor the entry of additional firms, which widens the range of varieties and increases consumers' utility who have a love for variety. As a consequence, the emergence of a large city is more likely to occur when the service sector is able to provide a larger number of differentiated varieties. Hence, labor-saving innovations such as the development of new information and communication technologies pushes toward the concentration of services in large cities.

By contrast, when $t/\delta > 2(1 + p^2)/\phi$, the symmetric equilibrium is stable. This is because the gains from variety do not compensate consumers for the higher urban costs they would bear in the large city. In this case, instead of seeking variety, consumers aim to reduce urban costs, and thus, the population is equally dispersed between the two cities. Dispersion may even take the concrete form of a larger number of smaller cities, which are determined by the trade-off between urban costs and the gains from variety. To sum up, when commuting costs steadily decrease, a service economy shifts from dispersion to agglomeration because the latter allows individuals to consume more services and to earn higher wages.

29.3.2 The Size and Industrial Structure of Cities

We now take a broader perspective by considering a two-sector economy in which labor is perfectly mobile between locations *and* sectors. The objective is to determine the interindustry distribution of consumers as well as their residential location between and within cities.

The economy involves a manufacturing sector supplying a freely tradable good ($\tau = 0$) and another sector producing a nontradable service (other than land) for local consumption. Focusing on such an industrial mix allows revisiting the export base theory grounded in the assumption that the urban economy can be divided into two very broad sectors, that is, a basic sector whose fortunes depend largely in external factors and a nonbasic sector which depends on local factors. The tenet of this theory holds that the basic sector is the prime cause of local economic growth (Tiebout 1956).

A. The Export Base Theory Revisited. The manufactured good is denoted by 1 and the nonhousing service by 2. The utility derived from consuming q_i units of a variety i of good $j = 1, 2$ is given by Eq. (29.1). In other words, the parameters associated with the utility arising from consuming one variety of the manufactured product or of the consumption service are identical. This assumption does not affect qualitatively the properties of the spatial equilibria. Indeed, because good 1-varieties are available everywhere at the same price, the consumer surplus generated by the consumption of the manufactured good is the same regardless of the city in which consumers live. Furthermore, the profits earned by the manufacturing firms are the same regardless of the city in which they are located. Thus, the equilibrium values of the consumer surplus and wage associated with

good 1 do not play any role in workers' decision to move. As a consequence, assuming that the parameters of Eq. (29.1) are the same for goods 1 and 2 entails no loss of generality for the determination of the sectoral and spatial structure of the economy.

Preferences now involve two nonhousing goods and are given by

$$U(q_0; q_{ij}) = \sum_{j=1,2} \left[\int_0^{n_j} q_{ij} di - \frac{\beta n_j}{2(n_1 + n_2)} \int_0^{n_j} q_{ij}^2 di \right. \\ \left. - \frac{1}{2(n_1 + n_2)} \int_0^{n_j} q_{ij} \left(\int_0^{n_j} q_{kj} dk \right) di \right] + q_0 \quad (29.14)$$

where q_{ij} is the quantity of variety $i \in [0, n_j]$ of good $j = 1, 2$. Because good 1 is tradable, the total number n_1 of good 1-varieties is available in both cities, whereas n_2 is the number of good 2-varieties supplied in the city where the consumer lives. Consumers having a love for variety may vary between goods and services; the second term of Eq. (29.14) is weighted by the ratio $n_j/(n_1 + n_2)$. This captures the idea that a good supplied as a small range of varieties has more impact on the consumer's well-being than a good made available through a large array of varieties. Note that the following analysis can be extended to cope with different attitudes toward variety by assuming that $\beta_1 \neq \beta_2$.

Let λ_{ir} be the number of consumers working in sector $i = 1, 2$ and living city $r = A, B$. Labor being mobile between cities and sectors, the variables λ_{ir} are endogenous and determined in equilibrium. Labor market clearing implies

$$n_1 = \frac{\lambda_{1A} + \lambda_{1B}}{\phi_1} \quad n_{2r} = \frac{\lambda_{2r}}{\phi_2} \quad (29.15)$$

Labor being mobile between sectors, in equilibrium, it must be that $w_r = w_{1r} = w_{2r}$. Letting $\lambda_r = \lambda_{1r} + \lambda_{2r}$ be the population residing in city r , the budget constraint of a consumer residing in city r may be written as follows:

$$n_1 p_1 q_{1r} + n_{2r} p_{2r} q_{2r} + \frac{t}{\delta} \frac{\lambda_r}{2} + q_0 = \bar{q}_0 + w_r$$

where p_1 is the common price of a good 1-variety, p_{2r} the consumer price of a good 2-variety in city r , and q_0 the consumption of the numéraire.

It is readily verified that the individual demand for a good i -variety in city r is given by

$$q_{1r} = \left(\frac{1}{1+\beta} - \frac{p_1}{\beta} + \frac{\bar{p}_1}{\beta(\beta+1)} \right) \left(1 + \frac{n_{2r}}{n_1} \right) \quad (29.16)$$

$$q_{2r} = \left(\frac{1}{1+\beta} - \frac{p_{2r}}{\beta} + \frac{\bar{p}_{2r}}{\beta(\beta+1)} \right) \left(1 + \frac{n_1}{n_{2r}} \right) \quad (29.17)$$

Whereas the average price \bar{p}_1 is defined over the entire range of good 1-varieties because good 1 is tradable, \bar{p}_{2r} is defined only over the range of good 2-varieties produced in city r . Although this demand system involves no income effect, it displays a rich pattern of substitution via the relative number of varieties. Specifically, when the number of good i -varieties available in city r increases, the individual demands for good j -varieties are shifted upward because good j becomes relatively more attractive. In particular, the size and distribution of the service sector (n_{2r}) affects individual demands for the manufactured good in each city (see Eq. (29.17)). Unlike the export base theory which maintains that the industries producing tradable goods are the economic base of the urban economy, the model used here shows that *a growing service sector impacts positively on the local demand for the tradable good*.

Likewise, the size of the manufacturing sector (n_1) affects the individual demand for services in each city and, therefore, the spatial distribution of this sector. In contrast, the distribution of manufacturing firms has no direct impact on individual demands for good 1 because trading this good is costless. This suggests that manufacturing firms are indifferent between locations. But they are not because their workers are attracted by cities supplying a wide range of services.

Let

$$\pi_{1r} \equiv p_1[q_{1A}(p_1)\lambda_A + q_{1B}(p_1)\lambda_B] - \phi_1 w_r$$

be the profits earned by a manufacturing firm established in city r . As in Sect. 29.2.1, when choosing its own price, each firm treats parametrically the wage w_r as well as the average prices \bar{p}_{1A} and \bar{p}_{1B} . Setting $\tau = 0$ in Eq. (29.7) yields the equilibrium price of good 1, which is constant and the same in both cities: $p_1^* = \mathbf{p}$. When they are not agglomerated, manufacturing firms therefore make the same operating profits in both cities. This implies that they pay the same wage w_1^* to their workers. As a consequence, there is factor price equalization: $w_1^* = w_{1r}^* = w_{2r}^*$. In this event, the urban cost differential is exactly compensated by the difference in the number of nontradable services supplied in each city. Simply put, *consumers choose to live in a larger city where they bear higher urban costs because they have access to a wider array of local services*. Profits being zero in equilibrium, the wage paid by a manufacturing firm is equal to

$$w_1^* = \mathbf{p}^2 \sum_{r=A,B} \frac{n_r \lambda_r}{n_1 \phi_1} \quad (29.18)$$

where $n_r = n_1 + n_{2r}$.

The profits made by a service firm set up in city r are given by

$$\pi_{2r} = p_{2r} q_{2r} \lambda_r - \phi_2 w_{2r}$$

where p_{2r} is the price quoted by such a firm. Because substitution effects go through the numbers of varieties only, the equilibrium price of a good 2-variety is given by

Eq. (29.13). This in turn implies that the equilibrium wage paid by the service firms located in city r is

$$w_{2r}^* = \mathbf{p}^2 \frac{n_r \lambda_r}{n_{2r} \phi_2} \quad (29.19)$$

which varies with the size (λ_r) and the sectoral mix (n_r/n_{2r}) of the city. Note that the service sector is never agglomerated. Otherwise, w_{2r}^* becomes arbitrarily large when there is no service firms in city r ($n_{2r} = 0$).

Since $S = 1$, the welfare of a consumer working in sector i and living in city r is given by

$$V_{ir} = n_1 + n_{2r} + w_r^* - \frac{t}{\delta} \frac{\lambda_r}{2}$$

This shows how consumers' well-being depends on the spatial and sectoral distribution of jobs.

B. Urban Hierarchy. A *spatial-sectoral equilibrium* arises when no worker has an incentive to change place and/or to switch job. The stability of such an equilibrium is studied by using the myopic evolutionary dynamics (Fujita et al. 1999):

$$\dot{\lambda}_{ir} = \lambda_{ir}(V_{ir} - \bar{V}) \quad (29.20)$$

where $\bar{V} = \sum_i \sum_r V_{ir}$ is the average utility in the entire economy. Note that in Eq. (29.20), the choices of jobs and locations are treated in a symmetric way.

In what follows, we focus on the case in which the manufacturing sector is not fully agglomerated ($0 < \lambda_{1r}^* < 1$). In this event, we have $w_1^* = w_{2r}^*$. Using $p_1^* = p_2^* = \mathbf{p}$, the wage equality implies that $\sum_r n_{2r}^* \phi_2 = n_1^* \phi_1$. As a consequence, the labor force is equally split between the two sectors ($\lambda_1^* = \lambda_2^* = 1/2$).

The utility differential is now given by

$$V_{ir} - \bar{V} = \frac{2\delta - \phi_2 t}{\phi_2 \delta} \left(\lambda_{2r} - \frac{1}{4} \right) \lambda_s - \frac{t}{\delta} \left(\lambda_{1r} - \frac{1}{4} \right) \lambda_s \quad (29.21)$$

where λ_s is the city s -population (recall that w_{ir}^* is equal to the average wage). Solving the system Eq. (29.21) shows that there are two candidate equilibria (up to a permutation between A and B):

$$\lambda_{1A}^* = \lambda_{2A}^* = 1/4 \quad (29.22)$$

and

$$\lambda_{1A}^* = \frac{1}{4} + \frac{(2\delta - \phi_2 t)\sqrt{\Delta}}{4\phi_1 \phi_2 t} \quad \lambda_{2A}^* = \frac{1}{4} + \frac{\sqrt{\Delta}}{4\phi_1} \quad (29.23)$$

where

$$\Delta \equiv \phi_1^2 + 2\phi_1 \phi_2 - 2\phi_1 \phi_2^2 t / \delta$$

In both configurations, the total number of good i -varieties is given by $n_i^* = 1/2\phi_i$, and thus, the industrial mix $n_1^*/n_2^* = \phi_2/\phi_1$ in the global economy depends on the relative productivity of labor in the two sectors. By contrast, when the asymmetric configuration prevails, *cities differ not only in size but also in industrial structures*.

As in Sect. 29.2, the symmetric pattern Eq. (29.22), which involves two cities having the same size and the same industrial mix, is always a spatial-sectoral equilibrium. On the other hand, the asymmetric configuration Eq. (29.23) is a (stable) equilibrium if and only if $\Delta > 0$, that is,

$$\frac{t}{\delta} < \frac{\phi_1}{2\phi_2^2} + \frac{1}{\phi_2}$$

In other words, commuting costs (population density) must be sufficiently low (high) for a large city (A) and a small city (B) to coexist. Moreover,

$$\lambda_A^* - \lambda_B^* = \frac{\delta\sqrt{\Delta}}{\phi_1\phi_2 t} \geq 0$$

implies that a lower t or a higher δ gradually enlarges the population gap between the two cities. Though workers are identical, there is no catastrophic bifurcation: Small changes in commuting costs generate small changes in the location *and* the composition of economic activities. In other words, accounting for the possibility of changing jobs smooths out the process of migration.

A few remarks are in order. First, the existence of a nontradable service selects a well-defined distribution of the footloose industry 1. More precisely, except for fairly high commuting costs, the nontradable sector acts as a centripetal force that results in the (partial) agglomeration of the manufacturing sector. Second, as long as $t < 2\delta/\phi_2$ holds, the larger city supplies a wider array of varieties of each good than the smaller city ($\lambda_{1A}^* > 1/4 > \lambda_{1B}^*$ and $\lambda_{2A}^* > 1/4 > \lambda_{2B}^*$). In this case, the urban system displays a Christaller-like *hierarchy*: By supplying a larger array of services, city A attracts more consumers than city B . Though the demand for the manufactured good is higher therein (see Eq. (29.16)), this does not attract more manufactured workers because this good is shipped at zero cost. Thus, the process of circular causation comes to an end. Note, however, that the population gap between the two cities grows when the service sector becomes more productive.

Third, when $t > 2\delta/\phi_2$ holds, the larger city has a larger labor share in the service sector, whereas the smaller city has a larger labor share in the manufacturing sector:

$$\frac{\lambda_{1B}^*}{\lambda_B^*} > \frac{1}{2} > \frac{\lambda_{1A}^*}{\lambda_A^*} \quad \frac{\lambda_{2A}^*}{\lambda_A^*} > \frac{1}{2} > \frac{\lambda_{2B}^*}{\lambda_B^*}$$

In this event, the urban system involves *diversified but relatively specialized cities*. This is because the size advantage associated with the larger city no longer

compensates enough manufacturing workers for the higher urban costs they would bear there. This result is consistent with Ricardo's comparative advantage theory: The larger city has a comparative advantage in nontradables because it has a larger local market; the smaller city's comparative advantage is its lower level of urban costs. Note that a city's comparative advantage is not given; it emerges from market interactions and labor mobility.

Fourth, and last, the export base theory predicts that an increase in the local size of the basic sector induces a more than proportionate increase in the city size. It is readily verified that the equilibrium condition $V_{iA}^* = V_{iB}^*$ yields

$$\lambda_A^* = \frac{1}{2} \frac{\delta - \phi_2 t}{2\delta - \phi_2 t} + \frac{2\delta}{2\delta - \phi_2 t} \lambda_{1A}^*$$

where $2\delta/(2\delta - t\phi_2) > 1$ is the "regional multiplier." Hence, a shock that makes the basic sector larger (λ_{1A}^*) boosts a more than proportionate growth of the city size (λ_A^*) by attracting more services. However, a larger nonbasic sector also leads to the expansion of the basic sector, which means that the nonbasic sector can be an engine for urban growth.

Observe that the impact of the nonbasic sector on total employment is higher in the larger city because the service sector is relatively more concentrated in city A than in city B when $t/\delta < 1/\phi_2$. Indeed, $V_{iA} = V_{iB}$ implies that

$$\lambda_A^* = \frac{\phi_2 t - \delta}{2\phi_2 t} + \frac{2\delta}{t\phi_2} \lambda_{2A}^*$$

so that the regional multiplier of the nonbasic sector exceeds of the regional multiplier of the basic sector when $t/\delta < 1/\phi_2$, an inequality which is more likely to hold when the productivity in the nonbasic sector is high.

29.4 The Future of Cities

One may wonder how the kind of approach surveyed in the foregoing sections may help understand some of the main challenges faced by cities in the twenty-first century. In what follows, we consider two different issues which have important policy implications: (i) the growing share of retirees in developed countries, whose income does not come from labor, and (ii) the environmental impact of the rapid urbanization in emerging countries like China and India.

29.4.1 Cities in Aging Nations

The old-age dependency ratio (the ratio people aged 65 and older to people aged 15 to 64) is projected to double by 2050 within the European Union, with four persons

of working age for every elderly citizen to only two. This ratio is expected to be lower in the United States, with a rise from 19 to 32 %, but higher in Japan, with a rise from 25 in 2000 to 72 % in 2050. Such demographic changes are likely to have a major impact on cities because the retirees are driven by location factors that differ from those governing workers' residential choices. Workers' welfare depends on local services, land rent, and wages, whereas rentiers' welfare depends only upon local services/amenities and land rent. As a consequence, when the share of old people takes on a sufficiently high value, the process of circular causation *à la Myrdal* sparked by workers' location choice could well be challenged.

To study how the urban system might change as the old-age dependency ratio rises, we consider the model of Sect. 29.3.2 in which the population is split between two groups of consumers, that is, the *elderly* and the *workers* whose respective numbers are $\rho \geq 0$ and $1 - \rho \geq 0$. City B is endowed with an amenity $a > 0$, which is valued only by the elderly. We close the model by assuming that land is collectively owned by the elderly. The income of a retiree is, therefore, given by the aggregate land rent (*ALR*) divided by the total number of elderly (ρ). Workers and retirees have different unit commuting costs, t and θ , respectively. We assume $\theta > t$. The case where $t > \theta$ leads to more cumbersome expressions which do not affect the nature of our main results. What matters for our purpose is that a city's urban costs increase with the number of retirees residing there.

Let s_r be the share of elderly people living in city $r = A, B$. City r -population is then given by $\lambda_r = (\lambda_{1r} + \lambda_{2r})(1 - \rho) + s_r \rho$. Besides λ_{1r} and λ_{2r} , we have to determine s_r . If the elderly are those living close to the CBD, workers' urban costs borne are now as follows:

$$UC_r = t \frac{(\lambda_{1r} + \lambda_{2r})(1 - \rho) + s_r \rho}{2\delta}$$

which is equal to Eq. (29.6) when $\rho = 0$. It thus varies with the distribution of activities as well as with the way the retirees distribute themselves between the two cities.

Because of the asymmetry in the amenity supply, the elderly's equilibrium condition is given by $V_A^o - V_B^o = a$ with $V_r^0 = n_1 + n_{2r} + ALR/\rho - UC_r^o$. The urban costs UC_r^o borne by the retirees are given by

$$UC_r^o = \theta \frac{s_r \rho}{2} + t \frac{(\lambda_{1r} + \lambda_{2r})(1 - \rho)}{2}$$

The equilibrium distribution of the elderly between cities is the same regardless of the spatial and sectoral allocation of workers:

$$s_B^* = \frac{1}{2} + \frac{a}{\rho(\theta - t)} \quad (29.24)$$

As expected, more elderly choose to live in the city endowed with the amenity advantage than in the working city. A larger share of elderly in the economy

increases the number ρs_B^* of old people living in city B . Likewise, the number of old people residing in city A increases, thus meaning that *the population of both cities gets older*.

When the share of the elderly people in the economy is not too high, the economy displays two stable equilibria. In the former one, the mobility of the elderly does not jeopardize the existing urban hierarchy, whereas it does in the latter one (Gaigné and Thisse 2009). This could explain why there are contradicting opinions regarding the evolution of urban systems in aging nations. The equilibrium in which the working city remains the primate city, while the other city accommodates the larger share of retirees, is the one that agrees with current empirical evidence (Chen and Rosenthal 2008). The corresponding equilibrium distribution of workers between sectors and cities is as follows:

$$\lambda_{1A}^* = \frac{1}{4} + \frac{(2 - t\phi_2)\sqrt{\Delta_a}}{4t\phi_1\phi_2} + \frac{a}{(1 - \rho)(\theta - t)} \quad \lambda_{2A}^* = \frac{1}{4} + \frac{\sqrt{\Delta_a}}{4\phi_1} \quad (29.25)$$

where

$$\Delta_a \equiv \phi_1(\phi_1 + 2\phi_2) - \frac{2t\phi_1\phi_2^2}{1 - \rho} < \Delta$$

Note that Eq. (29.25) boils down to Eq. (29.23) when $\rho = a = 0$.

Thus, *workers and retirees are not attracted by the same city*. Moreover, as shown by Eqs. (29.24) and (29.25), when city B 's local government improves its amenity supply, city B attracts a growing number of retirees, whereas the number of jobs in the working city rises. This provides a rationale for recent empirical evidence, which suggests that retirees and workers tend to live separately as the old-age dependency ratio increases.

Moreover, an aging population (a higher ρ) induces the dispersion of services at the expense of the working city while its effect on the manufacturing sector is ambiguous. In other words, an increasing share of retirees may challenge the performance of the working city. As a result, if the agglomeration of manufactures and services generates benefits not taken into account in the model, the economy will incur efficiency losses. In addition, employment in the working city decreases because the elderly city attracts more services. However, beyond some limit, the migration of retirees toward the amenity city raises the level of urban costs and/or decreases the supply of local services. This restores, to some extent, the attractiveness of the working city. Nevertheless, this need not be true for the services which still benefit from a big market in the elderly city. Regardless of old-age dependency ratio, the working city remains the larger one ($\lambda_A^* > \lambda_B^*$).

To sum up, though in an aging nation the relocation of consumption services weakens the supremacy of the working cities, these ones maintain their primacy. Indeed, as long as it is more profitable for the bulk of manufactures to congregate, a large share of services is prompted to set up therein. In addition, as the population gets older, cities diverge in their job and demographic structures. Yet, the supply of consumption services should prevent the complete spatial separation of workers and retirees.

29.4.2 Are Compact Cities Ecologically Desirable?

The transport sector is a large and growing emitter of greenhouse gases (GHG). It accounts for 30 % of total GHG emissions in the United States and approximately 20 % of GHG emissions in the EU-15. Moreover, road-based transport accounts for a very large share of GHG emissions generated by the transport sector. For example, in the US, nearly 60 % of GHG emissions stem from gasoline consumption for private vehicle use, while a share of 20 % is attributed to freight trucks, with an increase of 75 % from 1990 to 2006. Although new technological solutions will improve energy efficiency, other initiatives are needed, such as mitigation policies based on the reduction of average distances traveled by commodities and people.

A. The Ecological Trade-Off Between Commuting and Shipping Costs. We have seen that transporting people and commodities involves economic costs. It also implied ecological costs that obey the fundamental trade-off of Sect. 29.2.1: The agglomeration of firms and people in a few large cities minimizes the emissions of GHG stemming from shipping commodities, but increases those generated by longer commuting; dispersing people and firms across numerous small cities has the opposite costs and benefits. If cities are more compact (i.e., a higher population density δ), then, keeping population and firms fixed, the costs associated with the former spatial configuration (concentration) fall relative to those associated with the latter (dispersion) because people commute over shorter distances. However, when one recognizes that firms and people choose their location in order to maximize profits and utility, a policy that aims to make cities more compact will affect the intercity pattern of activities by fostering their progressive agglomeration, thus raising the level of GHG within fewer and larger cities. Therefore, the ecological effects of an increasing-density policy are *a priori* ambiguous (Gaigné et al. 2012).

To illustrate how this trade-off operates, we consider the model of Sect. 29.2.1 and assume that the carbon footprint E of the urban system stems from the total distance traveled by commuters within cities (C) and the total quantity of the manufactured good shipped between cities (T):

$$E = e_C C + e_T T$$

where e_C is the amount of GHG generated by one unit of distance traveled by a consumer, while shipping one unit of the manufactured good between cities generates e_T units of carbon dioxides.

Because consumers are symmetrically distributed on each side of the CBD, the value of C depends on the intercity distribution of the manufacturing sector and is given by

$$C = \frac{1}{4\delta} (\lambda_r^2 + \lambda_s^2)$$

Clearly, the emission of GHG stemming from commuting increases is minimized when the manufacturing sector is evenly dispersed between two cities ($\lambda_r = \lambda_s = 1/2$).

Regarding the value of T , it is given by the sum of equilibrium trade flows:

$$T = \frac{\phi[4\beta - (4\beta + 1)\tau]}{2(2\beta + 1)\beta} \lambda_r \lambda_s$$

where $T > 0$ because $\tau < \tau_{trade}$. As expected, T is minimized when consumers and firms are agglomerated within a single city ($\lambda_r = 0$ or 1). Note also that T increases when shipping goods becomes cheaper because there is more intercity trade. Hence, transportation policies that foster lower shipping costs give rise to a larger emission of GHG.

Thus, E is described by a concave or convex parabola in λ , so that the emission of GHG is minimized either at $\lambda = 1$ or at $\lambda = 1/2$. Therefore, it is sufficient to evaluate the sign of $E(1; \delta) - E(1/2; \delta)$, which is negative if and only if $\delta > \delta_e$ where

$$\delta_e \equiv \frac{e_C}{e_T} \frac{(2\beta + 1)\beta}{\phi[4\beta - (4\beta + 1)\tau]}$$

As a consequence, the agglomeration of activities within a single city is ecologically desirable if and only if $\delta > \delta_e$. Otherwise, dispersion is the best ecological outcome. As a consequence, *agglomeration or dispersion is not by itself the most preferable pattern from the ecological point of view*. Contrary to general beliefs, large compact cities need not imply low levels of pollution. For agglomeration to be ecologically desirable, the population density must be sufficiently high for the average commuting distance to be small enough.

B. Does the Market Yield a Good, or a Bad, Ecological Outcome? As seen in Sect. 29.2.1, $\lambda = 1/2$ is a stable equilibrium if δ is smaller than $\delta_m \equiv \phi t / (\tau)$. Otherwise, the manufacturing sector is concentrated into a single city. Because $\delta_m = 0$ at $t = 0$ and increases with t , while δ_e is independent of t , the two curves δ_m and δ_e intersect once. As a result, *the market yields either the best or the worst ecological outcome*.

Specifically, there exists a unique value \bar{t} such that $\delta_m = \delta_e$. Consider, first, the case where t exceeds \bar{t} . If $\delta < \delta_m$, the market outcome involves two cities. Keeping this configuration unchanged, a more compact city (i.e., a higher δ) always reduces the emissions of pollutants. Once δ exceeds δ_m , the economy gets agglomerated, thus leading to a downward jump in the GHG emissions. Further increases in δ allow for lower emissions of GHG. Hence, when commuting costs are high, a denser city always yields lower emissions of GHG. Assume now that $t < \bar{t}$. As in the foregoing, provided that $\delta < \delta_m$, the market outcome involves dispersion while the pollution level decreases when the city gets more compact. When δ crosses δ_m from below, the pollution now displays an upward jump. In other words, when commuting costs are low, *more compact cities need not be ecologically desirable*.

Consequently, once it is recognized that consumers and firms are mobile, what matters for the total emission of GHG is the mix between city compactness (δ) and city size (λ), thus pointing to the need of coordinating environmental policies at the local and global levels. In other words, environmental policies must focus on the urban system as a whole and not on individual cities.

When it is recognized that the internal structure of cities may change with population density (see Sect. 29.2.2), the ecological effects of an increasing-density policy are even more ambiguous: More compactness favors the centralization of jobs at the city center. Gaigné et al. (2012) point out that, unless commuting to SBDs generates a massive use of private cars, compact and monocentric cities may generate more pollution than polycentric and dispersed cities. By lowering urban costs without reducing the benefits generated by large urban agglomerations, the creation of SBCs would allow large cities both reducing GHG emissions and enjoying agglomeration economies.

29.5 Conclusions

The idea of spatial interaction is central to regional science. Broadly defined, spatial interaction refers to flows across space that are subject to various types of spatial frictions, such as traded goods, migrations, capital movements, interregional grants, remittances, and the interregional transmission of knowledge and business cycle effects. Though the NEG literature has for the most part focused on the mobility of goods and production factors, these issues are at the heart of NEG. Instead of writing one more review of the vast literature produced in the footsteps of Krugman (1991), we have chosen to highlight the role that NEG may play in understanding the process of urban development. Specifically, through several major trade-offs, we have covered a range of issues that highlight the working of urban systems. To do so, we have used very simple models, which vastly contrast with the heavy mathematical apparatus employed in the literature.

To a large extent, the lack of attention paid by economists to earlier contributions in regional science is unwarranted. Regional scientists and geographers have developed several models, such as those ranging from the entropy to the gravity and logit models, which have proven to be very effective in predicting and explaining different types of flows. By ignoring this body of research, economists have sometimes rediscovered the wheel and missed the opportunity of developing much earlier a sound theory of the space economy. But equally unwarranted is the acrimony expressed by many geographers soon after the diffusion of Krugman's work: They miss the importance of working with a fully consistent microeconomic model, especially the need of using a well-defined market structure and a precise specification of the externalities at work.

Cities of the twenty-first century face new and important challenges, such as climate change, aging population, crime, poverty, social exclusion, food security, the supply and management of transportation and communication infrastructure, and competition among the few world's largest cities. It is, therefore, fundamental to have sound theoretical models which can be used as guidelines in developing empirical research and designing new policies. Is NEG a useful tool? For many important urban questions, we believe the answer is yes. From the methodological standpoint, NEG has two major merits. First, the decisions made by firms and households are based on land rents, wages, and prices, which are themselves

endogenous and related to the size and structure of cities. Second, NEG takes into account the fact that households and firms may relocate between and within cities in response to major changes in their economic environment. NEG is connected with fast-growing economic fields that provide a set of tools and concepts, which can be used to tackle new and challenging issues.

Nevertheless, NEG suffers from a major drawback, which has been brushed aside in most of the literature: It is built on a two-location setting. Yet, it is well known that a firm's location is the balance of a system of forces pulling the firm in various directions. The new fundamental ingredient that a multilocation setting brings about is that spatial frictions between any two cities are likely to be different. As a consequence, the *relative position* of a city within the whole network of interactions matters (Behrens et al. 2007). Another key insight one can derive in a multilocation economy is that any change in the underlying parameters has in general complex impacts which vary in nontrivial ways with the properties of the graph representing the spatial economy. When there are only two locations, any change in structural parameters necessarily affects directly either one of the two cities or both. On the contrary, when there are more than two locations, any change in parameters that directly involves only two cities now generates spatial spillover effects that are unlikely to leave the remaining cities unaffected. More work is called for here but one should not expect a simple answer.

Last, the literature features two distinct models of competition in space (i.e., spatial competition à la Hotelling (1929) and monopolistic competition in Krugman-like settings). Each one seems to describe competition on two different spatial scales. Indeed, the former fits well competition "in the small," which involves shopping malls, retailers, and service providers located within the same city; the latter provides a fairly good approximation of competition "in the large," that is, competition among producers supplying several cities and countries. A theory encompassing both settings is needed to understand better how consumer prices are formed within different urban neighborhood as well as in cities having different sizes and morphologies. The industrial organization literature on vertical relationships linking upstream (global) and downstream (local) firms through carriers is a good point where to start.

References

- Anas A, Arnott R, Small KA (1998) Urban spatial structure. *J Econ Lit* 36:1426–1464
- Bairoch P (1985) De Jéricho à Mexico. Villes et économie dans l'histoire. Gallimard, Paris.
English translation: (1988) Cities and economic development: from the dawn of history to the present. University of Chicago Press, Chicago
- Baldwin RE, Forslid R, Martin P, Ottaviano GIP, Robert-Nicoud F (2003) Economic geography and public policy. Princeton University Press, Princeton
- Behrens K, Lamorgese AR, Ottaviano GIP, Tabuchi T (2007) Changes in transport and non-transport costs: local vs global impacts in a spatial network. *Reg Sci Urban Econ* 37:625–648
- Cavaillès J, Gaigné C, Tabuchi T, Thisse J-F (2007) Trade and the structure of cities. *J Urban Econ* 62:383–404

- Chen Y, Rosenthal SS (2008) Local amenities and life-cycle migration: do people move for jobs or fun? *J Urban Econ* 64:519–537
- Combes P-P, Mayer T, Thisse J-F (2008) Economic geography. The integration of regions and nations. Princeton University Press, Princeton
- Dixit AK, Stiglitz JE (1977) Monopolistic competition and optimum product diversity. *Am Econ Rev* 67:297–308
- Engel C, Rogers J (2001) Deviations from purchasing power parity: causes and welfare costs. *J Int Econ* 55:29–57
- Fujita M, Krugman P, Venables AJ (1999) The spatial economy cities, regions and international trade. The MIT Press, Cambridge, MA
- Gaigné C, Thisse J-F (2009) Aging nations and the future of cities. *J Reg Sci* 49:663–688
- Gaigné C, Riou S, Thisse J-F (2012) Are compact cities environmentally friendly? *J Urban Econ* 72:123–136
- Glaeser EL, Kolko J, Saiz A (2001) Consumer city. *J Econ Geogr* 1:27–50
- Handbury J, Weinstein D (2011) Is new economic geography right? Evidence from price data. NBER Working Paper No. 17067
- Helpman E (1998) The size of regions. In: Pines D, Sadka E, Zilcha I (eds) Topics in public economics. Theoretical and applied analysis. Cambridge University Press, Cambridge, pp 33–54
- Henderson V, Mitra A (1996) New urban landscape: developers and edge cities. *Reg Sci Urban Econ* 26:613–643
- Hicks JH (1969) A theory of economic history. Clarendon, Oxford
- Hotelling H (1929) Stability in competition. *Econ J* 39:41–57
- Jovanovic M (2009) Evolutionary economic geography. Location of production and the European Union. Routledge, London
- Krugman PR (1980) Scale economies, product differentiation, and the pattern of trade. *Am Econ Rev* 70:950–959
- Krugman PR (1991) Increasing returns and economic geography. *J Polit Econ* 99:483–499
- Myrdal G (1957) Economic theory and underdeveloped regions. Duckworth, London
- Ottaviano GIP, Thisse J-F (2002) Integration, agglomeration and the political economics of factor mobility. *J Public Econ* 83:429–456
- Ottaviano GIP, Thisse J-F (2005) New economic geography: what about the N? *Environ Plan A* 37:1707–1725
- Puga D (2010) The magnitude and causes of agglomeration economies. *J Reg Sci* 50:203–219
- Samuelson PA (1954) The transfer problem and transport cost. II. Analysis of effects of trade impediments. *Econ J* 64:264–289
- Tabuchi T, Thisse J-F (2002) Taste heterogeneity, labor mobility and economic geography. *J Dev Econ* 69:155–177
- Tabuchi T, Thisse J-F (2006) Regional specialization, urban hierarchy, and commuting costs. *Int Econ Rev* 47:1295–1317
- Tiebout CM (1956) Exports and regional growth. *J Polit Econ* 64:160–164
- Timothy D, Wheaton WC (2001) Intra-urban wage variation, employment location and commuting times. *J Urban Econ* 50:338–366

New Economic Geography: Endogenizing Location in an International Trade Model

30

Steven Brakman, Harry Garretsen, and Charles van Marrewijk

Contents

30.1	Introduction	570
30.2	Increasing Returns and Intra-industry Trade	571
30.2.1	Basic Ingredients of the Model	571
30.2.2	The Home Market Effect (HME) as a Volume Effect	573
30.2.3	The Home Market Effect as a Factor Price Effect	577
30.3	Adding Labor Mobility	578
30.3.1	Core NEG Model	578
30.3.2	Model Extensions	581
30.4	Empirical Testing	582
30.5	Policy Consequences	586
30.6	Conclusions	587
	References	587
	Further Reading	589

Abstract

In this chapter we first briefly discuss how the new economic geography literature (NEG) follows from and builds on international trade theory. We then turn to the main empirical implications of NEG. We highlight that the main problem

This chapter is partially based on earlier work by the authors. We do not give detailed references to our own work but readers interested in further details can consult our book on geographical economics (Brakman et al. 2009) for more extensive and detailed discussions and references, see in particular ▶ Chaps. 3, “Labor Market Theory and Models” and ▶ 8, “Land Use, Real Estate, and Housing Markets”. We also make use of Brakman and Garretsen (2009)

S. Brakman (✉) • H. Garretsen

Faculty of Economics and Business, University of Groningen, AV, Groningen, The Netherlands
e-mail: s.brakman@rug.nl; j.h.garretsen@rug.nl

C. van Marrewijk

Utrecht University School of Economics, University of Utrecht, TC, Utrecht, The Netherlands
e-mail: J.G.M.vanMarrewijk@uu.nl

with empirical applications of NEG is that a single test of the implications of the model combined is illusive because of the structure of the model. As a result the main consequences of the model are usually tested separately. And some of the implications of the model are also consistent with other models. We stress, therefore, that despite a real surge in empirical NEG inspired research, the empirical evidence is still rather sketchy and also that so far NEG-based policy advice is still mostly qualitative.

30.1 Introduction

The Nobel Prize committee that awarded the Nobel Prize in economics to Paul Krugman in 2008 stressed that the award was essentially given to him for his contributions in (mainly) three papers in two disciplines: international trade and economic geography (The prize committee of the *Royal Swedish Academy of Sciences* stated in its scientific background report (p.1): “*Traditionally, trade theory and economic geography evolved as separate subfields of economics. More recently, however, they have converged [to] become more and more united through new theoretical insights, which emphasize that the same basic forces simultaneously determine specialization across countries for a given international distribution of factors of production (trade theory) and the long-run location of those factors across countries (economic geography)*”). Krugman (1979) and Krugman (1980) deal with international trade (notably intra-industry trade), whereas Krugman (1991) extends the analysis of the first two papers by endogenizing the spatial allocation of economic activity. Both contributions became workhorse models for the two disciplines: the monopolistic competition model became the standard international trade reference for models incorporating intra-industry trade and the extension of this model, by allowing for factor mobility, became the core model of new economic geography (hereafter, NEG, also known as geographical economics).

In this chapter we will highlight the main characteristics of NEG, and in doing so we will not only explain the fundamentals of NEG and trace its origins to international trade theory but also mention some of the more recent developments. Next, we will illustrate the current state of affairs with respect to the empirical evidence for NEG and, related to this, the policy consequences of the model. Three features stand out. First, the combination of increasing returns to scale, imperfect competition and transport costs gives rise to the so-called home market effect. Second, the combination of the home market effect with interregional labor mobility endogenizes the location decisions of firms and footloose workers and hence the spatial allocation of both supply and demand. This setup allows for multiple equilibria, one of which is a core-periphery equilibrium. This explains why the model has also been used in urban economics to explain, for example, a system of cities. Third, despite a large and increasing literature on empirical evidence, a convincing test of NEG is still missing. This implies that policy advice based on the model should be handled with care and so far the basic policy contributions of NEG are of a qualitative nature.

In this chapter we will focus on the three aforementioned issues. A chapter like this is too short to provide a full survey, but the key issues will be introduced and explained (For extensive surveys or introductions to new trade theory, see, for instance, Feenstra (2004). For surveys of and introductions to the new economic geography, see the general references at the end of this chapter). In essence, by discussing the three topics, we will stress the most important contributions of NEG and explain the *tug of war* between the agglomeration and spreading forces that are active in the NEG models and their potential empirical and policy implications.

30.2 Increasing Returns and Intra-industry Trade

30.2.1 Basic Ingredients of the Model

During the 1970s, it became increasingly clear that the standard workhorse models of international trade were at odds with the facts. The Heckscher-Ohlin and the Ricardian model give a rationale for interindustry trade only. Empirical research (Grubel and Lloyd 1975), however, clearly showed that trade between (developed) countries was mainly in the form of intra-industry trade. The bulk of trade is in similar goods between similar countries, a puzzling phenomenon in neoclassical trade models. The theoretical challenge was to come up with a trade model that allowed for intra-industry trade. A possible explanation should center on the role of increasing returns to scale and on an imperfect competition in market structure. In Krugman (1979), a simplified version of the monopolistic competition model, as developed by Dixit and Stiglitz (1977), is introduced (see also Dixit and Norman 1980). The Dixit-Stiglitz model provides a fruitful way to model monopolistic competition. Almost instantly it became the preferred choice of researchers to model monopolistic competition, and it has become the benchmark model in various fields (see for a survey of contributions, Brakman and Heijdra 2004). We give a simplified version of the model below (The discussion in Sect. 30.2 is based on Brakman and GarreSEN (2009)).

30.2.1.1 Demand

Household utility is characterized by a love-of-variety effect that assumes that each variety c_i , $i = 1, \dots, n$, enters utility, U , symmetrically as an incomplete substitute; H is a homogeneous commodity which can serve as a numéraire; and M is often referred to as manufacturing:

$$U = H^{1-\delta} M^\delta, \text{ where } M = \left(\sum_{i=1}^n c_i^\rho \right)^{1/\rho} \text{ and } 0 < \rho \equiv \left(1 - \frac{1}{\varepsilon} \right) < 1 \quad (30.1)$$

where δ is the Cobb-Douglas share of M and $(1 - \delta)$ the Cobb-Douglas share of H and ρ is the elasticity of substitution between varieties. If the number of varieties is (very) large, firms consider the elasticity of demand ε as given. Utility maximization of Eq. (30.1) subject to the budget constraint gives (For a step by step

derivation of this two-stage maximization problem, see, for instance, Brakman, Garretsen, and Van Marrewijk (2009, Chap. 3))

$$c_i = \frac{p_i^{-\varepsilon} \delta w L}{\sum_j^n p_j^{1-\varepsilon}} \quad (30.2)$$

where p_i is the price of variety i and n the number of varieties. The term in the denominator is related to the price index for the manufactured goods. In what follows we assume that there is only one factor of production, labor L , with wage rate w .

30.2.1.2 Supply

For the explanation of intra-industry trade, it is necessary that similar goods are produced in different places. Intra-industry trade follows immediately in a multicountry setting if all varieties are consumed by all consumers (love-of-variety effect). A simple way to introduce (internal) economies of scale and ensuring that each variety i is produced by a single firm is via the following labor cost function:

$$l_i = \alpha + \beta x_i, \text{ where } \alpha, \beta > 0 \quad (30.3)$$

Labor, l_i , is the only production factor, which earns a wage w . The parameters α and β determine the fixed and marginal costs, αw and βw , respectively (the fixed costs give rise to the internal scale economies). Equation (30.3) implies that average costs are decreasing in the quantity of variety i that is produced, and this warrants that in the competitive equilibrium a particular variety is produced by the firm that had initially the largest market share and thus the lowest costs per unit of production.

The full-employment condition describes that the summation of Eq. (30.3) over all varieties equals total labor supply:

$$L = \sum_{i=1}^n l_i = \sum_{i=1}^n \alpha + \beta x_i \quad (30.4)$$

Firms are defined symmetrically which implies that:

$$p_i = p; x_i = x \quad (30.5)$$

for all i in equilibrium.

30.2.1.3 Equilibrium

The next step is to derive the market equilibrium. This gives the equilibrium output of each firm, x_i , the equilibrium number of varieties and hence the equilibrium number of firms, n , and it also yields the equilibrium price wage ratio, p_i/w_i . Profit maximization gives the familiar markup pricing rule, equating marginal costs to marginal revenue (dropping the index because of symmetry):

$$p = \frac{\varepsilon}{\varepsilon - 1} \beta w \text{ or } \frac{p}{w} = \frac{\varepsilon}{\varepsilon - 1} \beta \quad (30.6)$$

The zero profit condition implies that

$$0 = px - (\alpha + \beta x)w \Rightarrow \frac{p}{w} = \beta + \frac{\alpha}{x} = \beta + \frac{\alpha}{Lc_i} \quad (30.7)$$

Equations (30.6) and (30.7) together give the breakeven output, x , of a firm that is consistent with profit maximization and free entry and exit into the market: $x = \frac{(\varepsilon-1)\alpha}{\beta}$.

30.2.1.4 International Trade

The gains of international trade are present in the model outlined above but only in a rudimentary way. An increase in the available labor supply still shifts the average cost downward. This shift has implications for the number of varieties that are produced, which increases (see Eq. (30.4), $L/l_i = n$), but has no impact on other elements in the model. Consumers gain from trade because they consume more varieties than before international trade was allowed.

More interesting results can be derived by introducing transport costs. This is certainly true from a NEG perspective because the relevance of economic *geography* crucially hinges upon the presence of positive transport costs; without transport costs geography does not matter. The combination of increasing returns to scale (IRS) and transport costs implies that firms not only want to concentrate production in a single location (because of IRS) but they also care where in space they locate production (because of the transport costs). Firms prefer to locate where demand for the variety they produce is relatively large. This interplay between IRS, transport costs, and demand has become known as the home market effect, which is also the basis for NEG literature. Our discussion of the home market effect is in two parts: the more than proportional production of the increasing returns sector in the larger market (the volume effect, Sect. 30.2.2) and the higher wages of the increasing returns sector in the larger market (the price effect, Sect. 30.2.3). *The key issue is that with positive transport costs, the larger market offers location benefits that are absent in models that do not include transport costs.* We introduce this difference as the two versions have important consequences for empirical tests of the model that are not always taken into consideration.

30.2.2 The Home Market Effect (HME) as a Volume Effect

Iceberg transportation costs have the advantage that transportation costs can be introduced without having to deal with a transportation sector (For a critique of the iceberg depiction of transport costs, see Fingleton and McCann (2007)). Assume the iceberg costs are τ (with $\tau \geq 1$), that is, τ units have to be shipped in order for one unit to arrive in the other country. This raises the costs of imported varieties to $p\tau$.

Demand for a domestic variety now comes from two sources: domestic demand Eq. (30.8a) and foreign demand Eq. (30.8b). From Eq. (30.2), it are (where * indicates foreign variables)

$$x_i = \frac{p^{-\varepsilon}}{np^{1-\varepsilon} + n^*(\tau p)^{1-\varepsilon}} \delta w L \quad (30.8a)$$

$$x_i^* = \frac{(\tau p)^{-\varepsilon}}{n(\tau p)^{1-\varepsilon} + n^*(p)^{1-\varepsilon}} \delta w^* L^* \quad (30.8b)$$

Similar equations can be derived for the foreign country. From the discussion following Eqs. (30.6) and (30.7), we know that output per firm is fixed and equal to x in equilibrium. Goods market clearing in each country for the increasing returns sector gives, for the home country,

$$X \equiv nx = \frac{np^{-\varepsilon}}{np^{1-\varepsilon} + n^*(\tau p)^{1-\varepsilon}} \delta w L + \frac{n(\tau p)^{-\varepsilon}}{n(\tau p)^{1-\varepsilon} + n^*(p)^{1-\varepsilon}} \delta w^* L^* \tau \quad (30.9a)$$

and, for the foreign country,

$$X^* \equiv n^*x = \frac{n^*(\tau p)^{-\varepsilon}}{n(p)^{1-\varepsilon} + n^*(\tau p)^{1-\varepsilon}} \delta w L \tau + \frac{n^*p^{-\varepsilon}}{n(\tau p)^{1-\varepsilon} + n^*(p)^{1-\varepsilon}} \delta w^* L^* \quad (30.9b)$$

Note the additional τ multiplication terms in both expressions and also note that output level in both countries – for individual firms – is x . In Eq. (30.9a), part of the home exports to foreign melts during transportation, but it needs to be produced before it can melt, and similarly in Eq. (30.9b) for exports from foreign to home, hence the additional multiplication by τ .

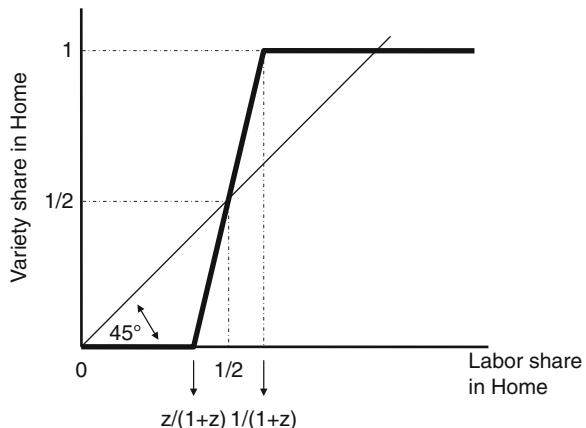
Assume first that there are no transport costs with respect to the homogeneous sector, H , and second (as is standard in international trade theory) that labor is mobile between sectors but *immobile* between countries. It follows that wages in the H sectors in both countries are identical and because of perfect inter-sector labor mobility, also in the increasing returns sector. Equation (30.6) allows us to choose units such that $p = w = 1$.

This implies that we can simplify Eqs. (30.9a)–(30.9b) as follows (with $Z \equiv \tau^{1-\varepsilon}$) (In the new economic geography literature, $\tau^{1-\varepsilon}$ is known as the freeness of trade; see Baldwin et al. (2003)):

$$\frac{x}{\delta} = \frac{1}{n + n^*Z} L + \frac{Z}{nZ + n^*} L^* \quad (30.9a')$$

$$\frac{x}{\delta} = \frac{Z}{n + n^*Z} L + \frac{1}{nZ + n^*} L^* \quad (30.9b')$$

Fig. 30.1 Home market effect, the volume effect



We have two equations and two unknowns, n and n^* . In principle we have three possible cases (numbered a to c), namely, complete specialization in one of the two countries (cases a and b) and incomplete specialization (case c):

- $n = 0, n^* = \frac{\delta(L+L^*)}{x}$, from Eq. (30.9b').
- $n = \frac{\delta(L+L^*)}{x}, n^* = 0$, from Eq. (30.9a').
- $n = \frac{\delta}{(1-Z)x}(L - ZL^*), n^* = \frac{\delta}{(1-Z)x}(L^* - ZL)$, from Eqs. (30.9a') and (30.9b').

Concentrating on the home country, we can distinguish between these three possibilities. If we introduce the following notation, $s_L = \frac{L}{L+L^*}$, $s_n = \frac{n}{n+n^*}$ where s_L is the labor share and s_n the share of varieties or firms in home, we arrive at

$$s_n = \begin{cases} 0, & \text{for } s_L \leq \frac{Z}{1+Z} \\ (1-Z)^{-1}[(1+Z)s_L - Z], & \text{for, } \frac{Z}{1+Z} < s_L < \frac{1}{1+Z} \\ 1, & \text{for, } s_L \geq \frac{1}{1+Z} \end{cases} \quad (30.10)$$

The first entry in Eq. (30.10) follows from combining case a with case c (where specialization of all increasing returns production in foreign just becomes binding). Similarly, the last entry follows from combining cases b and c. Finally, the middle entry follows from solving case number c. The implications become clear if we depict these three possible cases as in Fig. 30.1.

What we see is that if the home country is large (small) enough in terms of labor relative to foreign, it will attract (lose) all increasing returns manufactures. What is important in our discussion of the home market effect (HME) is the slope of the curve in the area $\frac{Z}{1+Z} < s_L < \frac{1}{1+Z}$. From Eq. (30.10), we know that the slope of the line-piece is $(1-Z)^{-1}(1+Z) > 1$, which implies that the larger country in this area has a *more than proportional* share of varieties and hence firms compared to its share in labor. The reasoning is as follows. Suppose that from the point $(\frac{1}{2}, \frac{1}{2})$

a foreign firm (together with its workers) relocates to the home country that now becomes the larger market (the reason why this might take place is unimportant). This increases the market by the amount of workers that move, but it also increases the spending power of existing consumers who no longer have to incur transport costs resulting from importing the variety. This “double” increase in demand raises profits in the larger market and attracts more firms to the increasing returns sector. Points on the solid line indicate that the increase in the number of firms must be more than proportional than the number of workers (some workers come from the homogeneous sector) in order to restore equilibrium.

Why do not all firms move to the larger market in order to restore equilibrium? The reason is that additional firms also introduce more competition that reduces the (potential) profits in the larger market. To explore the thought experiment of making the home market larger, it is instructive to look at the denominator of Eq. (30.8a). A firm moving from foreign to home makes the denominator smaller (as the variety no longer has to be imported), and this implies more local competition. This competition effect is stronger; the higher are transport costs (high transport costs shield a market from foreign competition). So fewer firms have to move to reestablish equilibrium following the movement of a firm from foreign to home if transport costs are high (the slope of the line gets closer to the 45° line).

To sum up, countries or regions with a relatively large demand for a good are home to a more than proportional share of production of that good. Against this home market or market size effect, the competition effect acts to ensure that in equilibrium, and depending on the model’s parameters (notably on the level of transport cost index, Z), not all firms in the differentiated IRS sector need to end up choosing the larger market as their location. From an empirical point of view, the model gives rise to a testable hypothesis with respect to international trade flows: countries with a relatively large home market for variety i ceteris paribus are net exporters of this variety. In the trade literature (see, e.g., Davis and Weinstein 2003), this implication of the home market effect has been subjected to a series of tests (see below).

Three other observations are relevant concerning the home market effect. The first one is that the effect is quite sensitive to the underlying assumptions. If international trade in the homogenous good is also subject to transport costs, the home market effect ceases to exist (Davis 1998). Also, the analysis of the home market effect quickly gets quite complicated (or even muddled) for the case of $n > 2$ regions or countries. The second observation is that in the example of Fig. 30.1, a large home demand (here, a large s_L) leads to an influx of firms where the necessary labor to enable the additional production has to be released from the homogenous sector. Given that international labor mobility is still impossible (as we will see the main difference between Krugman (1980) and Krugman (1991)), the additional demand for labor by the firms in the differentiated IRS sector in home does indeed fully materialize in higher production because of an infinitely elastic inter-sector labor supply.

If labor supply is not perfectly elastic, at least part of the response in the larger market will be in the form of higher wages (Fujita et al. 1999, Eq. (4.42);

Head and Mayer 2006). As we will see next, with a less than elastic labor supply, a relatively large demand or a larger home market then translates (partly) into higher wages. A third and final observation is that demand across locations is given. This is a direct consequence of the fact that workers and hence consumers are immobile between locations. Any demand or market size differences are therefore exogenously given. What happens if one drops this assumption? What if not only (IRS) firms but also (some) workers are mobile and can choose in which country or location they wish to live? Answering this question leads us to the center of NEG, but first we present another manifestation of home market effect in terms of wages. The reason is that migration is determined by (real) wage differences between locations. So, we first need to derive an expression for (real) wages.

30.2.3 The Home Market Effect as a Factor Price Effect

In the example underlying Fig. 30.1, we, by construction, ignored any effect that market or demand size differences might have on wages. Labor was perfectly elastic between sectors but not between countries, which is the usual assumption in international trade theory. This enables us to focus on the number of varieties (firms). In Krugman (1991), an opposing case is introduced; the larger market does not attract more than a proportional share of firms, compared to its share in labor, but all benefits of a larger market now show up in terms of higher wages in the increasing returns sector.

Actually such a wage effect can already be seen as an outcome of the Krugman (1980) model, we only have to change one assumption: labor is not only immobile between countries but now also immobile between sectors. The implications are that we no longer have factor price equalization and that the number of varieties (firms) is proportional to the given quantity of labor in the increasing returns sector (so by assumption the HME of the previous section is absent). The setup of the model remains the same, but we can no longer take the steps to simplify Eqs. (30.9a) and (30.9b) to Eqs. (30.9a') and (30.9b'). At the same time it is true that location in the larger market offers benefits relative to location in the smaller market. Again, as in the previous section, location in the larger market implies that firms do not have to incur transport costs and that this increases the spending (real income) of consumers. How does it show up in this case? We can use Eq. (30.9a) to show this for the home country [and similarly for the foreign country using Eq. (30.9b)]. Note that as wages are not necessarily the same, prices also differ between countries. Furthermore, we have to be careful how to define income, Y and Y^* , in this case; see below. Taking care of these aspects results in

$$\frac{(\varepsilon - 1)\alpha}{\beta} = \frac{np^{-\varepsilon}}{np^{1-\varepsilon} + n^*(\tau p^*)^{1-\varepsilon}} \delta Y + \frac{n(\tau p)^{-\varepsilon}}{n(\tau p)^{1-\varepsilon} + n^*(p^*)^{1-\varepsilon}} \delta Y^* \tau \quad (30.11)$$

In Eq. (30.11), we have again used the fact that in the model markup, pricing together with the zero profit condition fixes the break even output of firms [see the left-hand side of Eq. (30.11) and the discussion following Eqs. (30.6) and (30.7)].

Using $p = \frac{\varepsilon}{\varepsilon-1} \beta w$, and $p^* = \frac{\varepsilon}{\varepsilon-1} \beta w^*$, we can rewrite Eq. (30.11) in terms of wages in the manufacturing sector (and do the same for the foreign country):

$$w = \rho \beta^{-\rho} \left(\frac{\delta}{(\varepsilon-1)\alpha} \right)^{1/\varepsilon} \left(Y P_1^{\varepsilon-1} + \tau^{(1-\varepsilon)} Y^* P_2^{\varepsilon-1} \right)^{1/\varepsilon} \quad (30.12a)$$

$$w^* = \rho \beta^{-\rho} \left(\frac{\delta}{(\varepsilon-1)\alpha} \right)^{1/\varepsilon} \left(Y^* P_2^{\varepsilon-1} + \tau^{(1-\varepsilon)} Y P_1^{\varepsilon-1} \right)^{1/\varepsilon} \quad (30.12b)$$

where $P_1^{1-\varepsilon} = n(w/\rho)^{1-\varepsilon} + n^*(\tau w^*/\rho)^{1-\varepsilon}$, $P_2^{1-\varepsilon} = n(\tau w/\rho)^{1-\varepsilon} + n^*(w^*/\rho)^{1-\varepsilon}$, and Y and Y^* are the income generated in home and foreign, respectively.

These equations make sense in the following way. Wages in home are higher if it has a large home market in terms of real income, $Y P_1$, or if it is located near a large foreign market (large $Y^* P_2$ and low transport costs or, equivalently, a high freeness of trade, $\tau^{1-\varepsilon}$). The benefits of a large market are not reflected in a more than proportional share of firms relative to the labor share but in higher wages.

30.3 Adding Labor Mobility

30.3.1 Core NEG Model

It is now only a small step to make the model a full general equilibrium model that includes labor mobility (See Brakman and Garretsen (2009)). The only thing to add is the possibility of labor migration between regions. This implies that a region's market size becomes endogenous when migration is allowed to take place. In the 2-region setting of Krugman (1991), the equilibrium conditions of the model can be stated as follows:

$$Y = wL + 0.5L_H \quad (30.13a)$$

$$Y^* = w^*L^* + 0.5L_H \quad (30.13b)$$

$$w = \rho \beta^{-\rho} \left(\frac{\delta}{(\varepsilon-1)\alpha} \right)^{1/\varepsilon} \left(Y P_1^{\varepsilon-1} + \tau^{(1-\varepsilon)} Y^* P_2^{\varepsilon-1} \right)^{1/\varepsilon} \quad (30.13c)$$

$$w^* = \rho \beta^{-\rho} \left(\frac{\delta}{(\varepsilon-1)\alpha} \right)^{1/\varepsilon} \left(Y^* P_2^{\varepsilon-1} + \tau^{(1-\varepsilon)} Y P_1^{\varepsilon-1} \right)^{1/\varepsilon} \quad (30.13d)$$

$$P_1^{1-\varepsilon} = n(w/\rho)^{1-\varepsilon} + n^*(\tau w^*/\rho)^{1-\varepsilon} \quad (30.13e)$$

$$P_2^{1-\varepsilon} = n(\tau w/\rho)^{1-\varepsilon} + n^*(w^*/\rho)^{1-\varepsilon} \quad (30.13f)$$

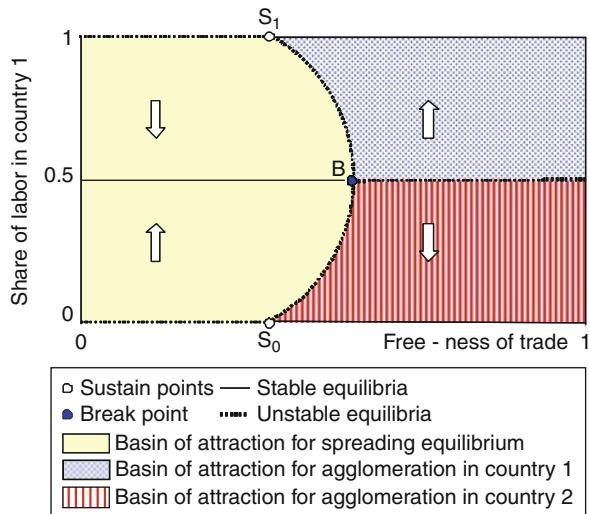
$$\omega = \frac{w}{P_1^\delta}, \quad \omega^* = \frac{w^*}{P_2^\delta} \quad (30.13g)$$

$$\frac{dL}{L} = -\frac{dL^*}{L^*} = \eta(\omega - \varpi), \quad \text{with } \varpi = \lambda\omega + \lambda^*\omega^* \quad (30.13h)$$

The model clearly builds (and even largely overlaps) with the international trade model of Sect. 30.2 but also includes some new elements of which interregional labor mobility is the most relevant one. Equations (30.13a) and (30.13b) are the income equations in the two regions or countries, home and foreign. The first term on the right-hand side indicates income earned in the increasing returns sectors that earn wages w and w^* in home and foreign, respectively. We assume that labor (in the increasing returns sector) is mobile between countries but not between sectors. The distribution of labor in the homogeneous (agricultural) sector is given and does not change. Total labor supply in this sector is L_H , and we assume – just for simplicity – that it is equally distributed over the two countries. There are no transport costs in this sector implying that wages earned in the homogeneous goods sector are equal in both regions, and we can use this sector as the numeraire sector, and wages in the increasing returns sector are relative to the wages in the homogeneous goods sector. It is important to note that we cannot do without this homogeneous goods sector in Krugman's (1991) core NEG model. It implies that even when labor in the increasing returns sector is completely agglomerated by being located in just one of the two regions, there is always a positive (residual) demand in the other region, and firms might want to relocate to this region in order to get away from the stiffer competition in the larger region.

Equations (30.13c)–(30.13f) are already familiar from earlier sections. Equations (30.13g) and (30.13h) give the dynamics in the model. Next, we define real income of a worker in the IRS sector in Eq. (30.13g). It is simply wages divided by the price index of all the commodities consumed (including the homogenous good). As the increasing returns to scale sector comprises a share δ in the consumption basket, we want to correct for this in Eq. (30.13g) (Note that P_1 and P_2 are price indices associated with the CES sub-utility indices, which explains the somewhat complicated notation of these expressions; see Brakman, Garretsen, and Van Marrewijk (2009, Chap. 3) for a detailed discussion of these price indices). We also divide by the price in the homogeneous sector (raised to the power $1-\delta$, the share of the homogeneous goods sector), but this does not show up because the homogeneous good is the numéraire good (and the price equals one). Equation (30.13h) states that labor in the increasing returns sectors moves to the region with the highest real wage. Of course, in the real world migration, decisions are based on much more than just real wage differences. The model easily gets quite complicated because if labor moves, to say, the home country, this changes incomes [Eqs. (30.13a) and (30.13b)] which affects nominal wages

Fig. 30.2 The Tomahawk for the Krugman (1991) model



[Eqs. (30.13c) and (30.13d)] and also the price indices [Eqs. (30.13e) and (30.13f)], which subsequently affect the migration decision itself, and given the functional forms of the model, these effects are nonlinear.

Given the key model parameters, most importantly the value of transport costs, the balance between the agglomeration forces (home market effect, price index effect) and the spreading forces (competition effect) determines what the equilibrium spatial allocation will be. It turns out that the model has basically three (stable) equilibria: full agglomeration in home or foreign and perfect spreading. Interestingly, the model is characterized not only by multiple equilibria but also by path dependency. Figure 30.2 illustrates the model. The so-called Tomahawk depicted in Fig. 30.2 shows that for low *freeness of trade* $\tau^{1-\epsilon}$ (=Z in the previous section), that is, for high transport costs τ , footloose labor is evenly spread between the two regions, but if the freeness of trade gets high enough, that is, if transport costs get low enough, all footloose workers end up in either region one or two in equilibrium.

The solid lines indicate stable equilibria and the dashed lines indicate unstable equilibria. The arrows indicate in what direction the incentive for firms (and footloose labor) points, depending on the value of transportation costs.

What are the forces that determine interregional migration? Three forces matter in the Krugman (1991) model: the price index effect, the home market effect, and the extent of competition effect. The price index effect stimulates agglomeration in the larger market as fewer varieties have to be imported, and this saves on transport costs. This effect is magnified by the home market effect discussed above. If the home market effect results in higher wages (see Sect. 30.3.1), it makes the larger market more attractive. These agglomeration effects are counteracted and diminished by the extent of the competition effect, which acts as the spreading force.

If a firm moves to the larger market, the denominators in Eqs. (30.9a) and (30.9b) become smaller, which reduces the demand for an individual firm. The more firms (and workers) there are in a region, the higher the level of competition will be.

The balance between these three forces determines the direction of the arrows in Fig. 30.2. For low values of transport costs (high values of the freeness of trade), this competition effect is felt less as the price difference between markets becomes smaller. Note from Fig. 30.2 that there is not a gradual change from one stable equilibrium to another but instead a catastrophic change; the moment the balance tilts between these forces, it is either full agglomeration in one region or the other. Starting from an initial situation of a low freeness of trade (left part of x-axis in Fig. 30.2), the point at which this happens is the so-called *break point*, B; moving from high to low transportation costs, spreading is no longer a stable equilibrium (breaks) if transport costs are reduced further. One could also start with very low transport costs (high freeness of trade) and then subsequently increase transport costs (lower the freeness of trade) until agglomeration becomes unstable. This happens at the so-called *sustain points*, S, in Fig. 30.2. (Note that in the middle part of Fig. 30.2, there is some overlap as to the range of the freeness of trade for the agglomeration and spreading equilibrium which indicates that the model is characterized by path dependency, see Brakman, Garretsen, and van Marrewijk (2009, Chap. 4) for an explanation).

30.3.2 Model Extensions

The model described in Sect. 30.3.1 states the essence of the NEG model that was introduced by Krugman (1991). In the subsequent literature, many additions have been incorporated in this model. These extensions are often motivated to correct some of the more unlikely aspects of the model or to make the model more tractable.

- The model introduced above can be extended with an intermediate production sector. Assuming that labor is intraregionally mobile but interregionally immobile produces more realistic results than the extreme outcome as described in Fig. 30.2. Economic integration in this case results in real wage convergence between regions rather than divergence (The reason for this is that the peripheral region becomes more attractive for manufacturing production as transport cost decline, because wage differences start to dominate transportation costs [which decline during economic integration]). Most importantly, however, is the extension introduced by Puga (1999). The model introduced above predicts that small changes in the parameter values could result in sudden and dramatic changes (see the sustain and break points in Fig. 30.2) which seems unrealistic in practice. Puga (1999) extends the intermediate production model and assumes that the numeraire sector is no longer characterized by constant returns to scale but instead by diminishing returns. This implies that pulling workers out of the homogeneous sector raises marginal productivity and nominal wages in

this sector. This adds an additional spreading force into the model preventing a bang-bang solution as in the standard model. Puga (1999) shows that this additional force combined with the assumptions in Krugman and Venables (1995) and Venables (1996) is so strong that instead of the Tomahawk as depicted by Fig. 30.2, a bell-shaped curve appears, which suggests a more gradual change from full agglomeration to complete spreading. This aspect of the Puga (1999) model makes it a preferred model in empirical research (see below).

- Another extension is to allow for more factors of production. One can assume, for instance, that manufacturing production uses high-skilled and low-skilled labor in production. This makes the model more realistic. A surprising side effect is that the model becomes more tractable than in the standard case. If high-skilled labor is used in the fixed part of production (α , in Eq. (30.3)) and low-skilled labor in variable-cost part of manufacturing production (β in Eq. (30.3)), but also in the production of the homogeneous sector, we no longer have to solve for nominal wages (which also can be normalized) but only for high-skilled wages. The solution is relatively straightforward. So, besides introducing the more realistic assumption of more factors of production, we can now derive explicit solutions. This is not the only change to the basic model that results in analytical solutions. Ottaviano, Tabuchi, and Thisse (2002), for instance, drop the CES demand structure and introduce a quasi-linear demand structure.
- The Melitz (2003) revolution has also entered NEG. Firms in the standard models are all the same according to their cost structure, see Eq. (30.3). However, Baldwin and Okubo (2006) introduce productivity differences between firms. They show that firms line up for reallocation from a small to a large market in order of firm productivity levels; more productive firms can already relocate at transportation levels that would imply a loss for less productive firms. Models like these are important for empirical research as they point toward an empirical complication: do larger markets benefit from “agglomeration economies” or from the fact that they are home to the more productive firms?

30.4 Empirical Testing

Can we test the main implications of NEG? This seems a simple question, but it has turned out that this question is surprisingly difficult to answer. The model has interesting consequences, but a combined test of the main or, let alone, all aspects is still missing. Head and Mayer (2004, p.2616) identify five main characteristics – slightly restated by us below and compressed into three main testable implications– that are special for NEG and could be tested to explain the facts implied by Figs. 30.1 and 30.2 (Brakman et al. 2009, Chap. 5):

- (a) *The home market effect:* Large regions will be home to a disproportional share of the imperfectly competitive industry. Such large markets are, therefore, net exporters of industries characterized by increasing returns to scale. As we discussed at some length in Sect. 30.2, there are also two other possible testable implications from this effect:

- (a1) The *volume version* (recall Fig. 30.1): A large market potential induces factor inflows from the small to the large market. Footloose factors of production will be attracted to those markets that pay relatively high real factor rewards. This leads to a process of circular causality.
- (a2) The *factor price version* [Eqs. (30.12a)–(30.12b)]: A large market potential raises local factor prices in the core relative to the periphery. An attractive market with a strong home market effect will increase demand for factors of production, and this raises factor rewards.
- (b) *Shock sensitivity*: As we discussed with Fig. 30.2, changes in the economic environment can trigger drastic and permanent changes in the spatial distribution of economic activity.
- (c) At some critical level of transport or trade costs, and again see Fig. 30.2, a further reduction in transport costs induces agglomeration by relocation of the footloose factors of production. This implies that more economic integration should at some point lead to (more) agglomeration of the footloose activities and factors of production.

Characteristics (a1) and (a2) describe the consequence for factors of production or factor prices once the home market effect is established. As is explained in Fujita et al. (1999, p. 57), the equilibrium of the Krugman (1991) model implies the following equation: $\frac{dY}{Y} = \gamma_1 \frac{dw}{w} + \gamma_2 \frac{dL}{L}$, where Y is total demand for the footloose sector, w is the nominal wage rate in this sector, and L is employment in this sector and γ_1 and γ_2 are parameters (This equation also shows why the findings on the HME show a highly variable pattern of estimated coefficients: both wages and employment changes should be accounted for, not only employment changes as in the strict version of the home market effect). It shows that an increase in the demand (Y) for the goods from the footloose sector not only causes employment changes (the volume version of the home market effect) but also induces wage (w) changes (the factor price version of the home market effect).

In a series of papers, Davis and Weinstein (2003) have developed an empirical methodology that enables them “*to distinguish a world in which trade arises due to increasing returns as opposed to comparative advantage*” (Davis and Weinstein 2003, p. 3). In general they find some support for the volume version of the home market effect. As shown in Brakman, Garretsen, and Schramm (2006), however, both effects are typically at work. On balance it appears that the wage channel is the main route toward spatial equilibrium (Head and Mayer 2006). This explains also why most empirical work has focused on the wage equations, Eqs. (30.13c)–(30.13d). Despite the empirical evidence that supports the volume or factor price version of the home market effect, the question remains whether this evidence is a test of NEG as such; they are also a characteristic of standard trade models. We will return to this question at the end of the present section.

One of the key elements of NEG is the shock sensitivity. As illustrated by Fig. 30.2, small changes in parameters (in casu, the level of transport costs) can (but need not) have big consequences. It implies for instance that a small change in economic integration could lead to spectacular changes in the spatial

distribution of economic activity. If small changes already can have large effects, one would be inclined to think that permanent effects in the spatial distribution of economic activity can be found after large changes. The key issue is whether one can come up with real world examples of a large, temporary, and exogenous shock that can act as a testing ground for the shock sensitivity hypothesis. In a seminal paper, Davis and Weinstein (2002) use the case of the allied bombing of Japanese cities during World War II (WW II) as an example of such a shock. Brakman, Garretsen, and Schramm (2004) apply the Davis and Weinstein (2002) approach to the case of the allied bombing of German cities during WW II. In both studies the question is the same: did individual cities return to their initial, prewar growth path after WW II? The breakup of Germany in 1949 in the Federal Republic of Germany FRG (West Germany) and the German Democratic Republic GDR (East Germany) and the subsequent reunification of the two Germanies in 1990 after the fall of the Berlin Wall is another example of a large, temporary (40 years) shock. Redding and Sturm (2008) use this shock to test whether *West-German* border cities (close to the FRG-GDR border) experienced a substantial decline compared to non-border cities in West Germany. The evidence of these studies is somewhat mixed. Davis and Weinstein (2002) do not find evidence of long term effects, whereas the other studies – on Germany – do find such effects. In general it seems that economies show some shock sensitivity. Again, and notwithstanding this evidence on shock sensitivity, the ultimate question for our present purposes is whether these studies do provide a real test of NEG as such. Note for instance that the NEG model as depicted by Fig. 30.2 allows for shocks that can have permanent and nonpermanent effects. Also, it is clear that NEG is not the only location or spatial model to predict that shocks can alter the spatial equilibrium allocation of economic activity.

Finally, NEG models predict that changes in transportation costs could result in changes in the degree of agglomeration through the relocation of the mobile factors of production. To this end, we essentially need the full model, as described in Sect. 30.3. The long run equilibrium equation relates migration to real wage differences, which are determined in the model. For empirical research this is a challenging consequence of the model. First of all we have to find out where we are in terms of for instance Fig. 30.2. Is the economy that we are looking at initially to the left or to the right of the break point? This is important, because we like to know what happens if transportation costs change. In real world applications, however, we deal with a multi-region world and implicitly confront this multi-region model with break points from the Tomahawk diagram in a 2-region setting, which is problematic. Similar analytical solutions for break points in a multi-region setting ($n > 2$) only exist if all regions are at equal distance from each other. This assumption effectively means that the actual geography (where regions are located on a map) does not play a role and thus that space is neutral in that sense. Any real world application clearly violates this assumption.

How should we proceed to arrive at more conclusive evidence for our empirical hypothesis on transportation or trade cost-induced agglomeration? One option is to

drop the 2-region model, with its analytical solutions, and instead use multi-region model simulations in which key equations are based on multi-regional estimates of Eqs. (30.13c)–(30.13d). In Bosker et al. (2010), this is the preferred option. They show that, in a qualitative sense, the multi-region non-neutral space model gives rise to the same conclusions as the simple 2-region version of the Puga (1999) model, that is, the results show that the 2-region model carries over in the multi-region case. Given this result, the answer to the question “where on these curves are we?” can be answered in a simulation setting. Repeated simulations, using the estimated wage curves for different values of transportation costs, allow us to construct a multi-region version of Fig. 30.2. Confronting this curve with actual estimates for transportation costs gives us an idea where on the curve we are, and in what direction the economy is moving, toward further agglomeration or further spreading. But it is clear that evidence based on simulations is not the same thing as evidence based on actual estimations, while using the structural equations of the model and using real data, of the third NEG hypothesis.

How convincing is the empirical evidence for the NEG studies related to the three empirical hypotheses that were outlined at the beginning of this section? In addition to the comments made above for each of the hypotheses, four general problems for empirical confirmation of the geographical economics model stand out (see also Redding 2010):

- (i) Studies are not only consistent with geographical economics models but also with other theories of trade and location (the home market effect can, for instance, also be found in other trade models; see also discussion in Fingleton and Fischer (2010)).
- (ii) Applying two region NEG models like Krugman (1991) to a multi-region world makes conclusive testing difficult or even outright impossible.
- (iii) Causality: Are the empirical observations caused by NEG forces or not?
The empirical evidence indicates that wages (left-hand side of the empirical specification of Eqs. (30.13c)–(30.13d)) are related to measures of market access. An important problem is whether this is a causal relation. Higher wages in regions with good market access may be caused by better institutions in surrounding regions or locational fundamentals instead of NEG forces, and the measures of market access might simply capture these more fundamental causes. This issue is more problematic for testing the home market effect (hypothesis 1) than for shock sensitivity tests (hypothesis 2, where cause and effect are more clearly distinguished).
- (iv) Using micro data: Virtually all empirical NEG work is based on the representative firm and consumer framework and ignores the extensive micro data sets that have become available over the past years. Using these data (as in the urban economics literature) may make it possible to determine if the agglomeration effects in the core are based on selection effects (truncation of the distribution) or agglomeration as such (rightward shift of the distribution); see Combes et al. (2008).

30.5 Policy Consequences

The NEG framework is widely used to discuss policy implications of (local or national) interventions. This holds, for example, for many (regional) studies performed on behalf of the European Union and for the recent *World Development Report 2009* by the World Bank (2008). A good summary of the six general policy conclusions based on the NEG model is provided by Ottaviano (2003); for a more extensive treatment of these points, see Baldwin et al. (2003):

- *Regional side effects.* One of the fundamental insights from NEG is that regions are connected and cannot be studied in isolation. Regional policy measures that, for example, affect economic integration, have consequences for all regions, not only the region at which the measure is aimed. Effects of regional policies are economy-wide.
- *Trade interaction effects.* Outcomes of the model depend crucially on initial levels of economic integration. A similar policy measure can have different effects, depending on the initial position of the economy (see Fig. 30.2).
- *Lock-in effects.* Temporary measures can have permanent effects. Suppose that a temporary subsidy takes an economy over the break point in Fig. 30.2. It is then possible that a new long run stable equilibrium is reached. The economy will remain there even when the subsidy is ended. Also history matters. If an economy finds itself in a stable equilibrium, strong policy measures might be needed in order to establish another equilibrium, in other words “history matters.”
- *Selection effects.* Figure 30.2 indicates that if transport costs are low, two stable equilibria are possible. Selecting one of the possibilities can have huge consequences from a welfare perspective. For example, the immobile workers in the core region benefit from being located in the core, but policy makers have to make up their mind that they indeed give welfare in the core region a greater weight in social welfare considerations than in the peripheral regions.
- *Threshold effects.* Policies measures can seem ineffective. The reason is that measures should take an economy over the break point in order to become effective.
- *Coordination effects.* NEG models are characterized by multiple equilibria. Especially in the overlap area in Fig. 30.2, expectations about the future of the economy can be important. If policy makers can convince firms/workers to relocate, this will start a self-sustaining move to a new equilibrium. A subsidy might take an economy toward a new equilibrium, but if policy makers can convince workers and firms that a specific region is the place-to-be, a subsidy is not required.

The list suggests that a world characterized by NEG offers policy makers many attractive options. However, some qualifications are in order. First of all which NEG model describes the world best, the stylized model depicted in Fig. 30.2 or one of the models that are extensions of this Fig. 30.2 model? The model in Fig. 30.2 is to a large extent driven by a few parameters, and it is highly unlikely that the real world can be described by only those few parameters, which are most likely

different for all sorts of economies or periods (Combes 2011). Furthermore, in general, core regions are always better off than peripheral regions. The Tomahawk Fig. 30.2 suggests that it is always possible, with the right measure, to pick a preferred equilibrium. Neary (2001) strongly argues against this “picking equilibria” role for the government, because, as we concluded at the end of the previous section, the empirical evidence for NEG is still too weak (see above), and that such a policy would bear the risk of strategic, and wasteful, rent seeking behavior from competing regions. Still, we think that one very strong policy measure results from NEG; regions are not free-floating islands in space, but they are spatially interdependent. All too often regional policies are addressed to deal with a specific regional issue – like low wages and lack of employment – and deal with such a region as if it is an island in space. Policy measures can have unexpected results. An investment in, for example, the regional infrastructure in this peripheral region might not stimulate growth in the periphery, but instead might strengthen the position of the core region because economic integration further strengthens the position of the core (see Fig. 30.2). This is probably – in a qualitative sense – one of the most important policy conclusions that can be derived from NEG.

30.6 Conclusions

We briefly discussed the structure of the NEG models to argue that the new aspect of this type of model is endogenizing economic size of a location. The economic aspect of the name refers to the economic tools used, while the geography part focuses on the crucial role of spatial interdependencies through transport and interaction costs. We then turn to the main empirical implications of NEG as summarized in a number of empirical characteristics. Despite the surge in empirical research in this area in the last decade, a number of crucial problems with empirically testing the NEG models remain. We list four of these, namely, (i) some effects can also be explained by other models, (ii) most tests in a multi-region world are only loosely based on a two-region basic model, (iii) causality problems are rarely adequately addressed, and (iv) we need to integrate locational phenomena at different scales by also using micro data. In view of our discussion of these shortcomings, the main policy implications (as discussed in Sect. 30.5) are still mostly qualitative, thus lacking a solid quantitative basis in most applied work.

References

- Baldwin R, Okubo T (2006) Heterogeneous firms, agglomeration and economic geography: spatial selection and sorting. *J Econ Geogr* 6(3):323–346
- Bosker M, Brakman S, Garretsen H, Schramm M (2010) Adding geography to the new economic geography; bridging the gap between theory and empirics. *J Econ Geogr* 10(6):793–823
- Brakman S, Garretsen H, Schramm M (2004) The strategic bombing of German cities during WWII and its impact on city growth. *J Econ Geogr* 4:201–218

- Brakman S, Heijdra BJ (eds) (2004) *The monopolistic competition revolution in retrospect*. Cambridge University Press, Cambridge
- Brakman S, Garretsen H, Schramm M (2006) Putting new economic geography to the test: free-ness of trade and agglomeration in the EU regions. *Reg Sci Urban Econ* 36(5):613–636
- Brakman S, Garretsen H (2009) Trade and geography: Paul Krugman and the 2008 Nobel prize in economics. *Spat Econ Anal* 4(1):5–23
- Brakman S, Garretsen H, Van Marrewijk C (2009) *The new introduction to geographical economics*. Cambridge University Press, Cambridge
- Combes P-P (2011) The empirics of economic geography: how to draw policy implications? *Rev World Econ* 147(3):567–592
- Combes P-P, Duranton G, Gobillon L (2008a) Spatial wage disparities: sorting matters! *J Urban Econ* 63(2):723–742
- Davis D (1998) The home market, trade, and industrial structure. *Am Econ Rev* 88:1264–1277
- Davis DR, Weinstein DE (2002) Bones, bombs and breakpoints: the geography of economic activity. *Am Econ Rev* 92(5):1269–89
- Davis DR, Weinstein DE (2003) Market access. Economic geography and comparative advantage: an empirical assessment. *J Int Econ* 59(1):1–23
- Dixit A, Norman V (1980) *Theory of international trade*. Cambridge University Press, Cambridge, UK
- Dixit A, Stiglitz J (1977) Monopolistic competition and optimum product diversity. *Am Econ Rev* 67(3):297–308
- Feenstra RC (2004) *Advanced international trade: theory and evidence*. Princeton University Press, Princeton
- Fingleton B, McCann P (2007) Sinking the iceberg? On the treatment of transport costs in new economic geography. In: Fingleton B (ed) *New directions in economic geography*. Edward Elgar, Cheltenham, pp 168–204
- Fingleton B, Fischer MM (2010) Neoclassical theory versus new economic geography: competing explanations of cross-regional variations in economic development. *Ann Reg Sci* 44(3):467–491
- Fujita M, Krugman PR, Venables AJ (1999) *The spatial economy; cities, regions, and international trade*. MIT Press, Cambridge
- Grubel HG, Lloyd P (1975) *Intra-industry trade: the theory and measurement of international trade in differentiated products*. Macmillan, London
- Head K, Mayer TH (2006) Regional wage and employment responses to market potential in the EU. *Reg Sci Urban Econ* 36(5):573–594
- Krugman P, Venables A (1995) Globalization and the inequality of nations. *Quart J Econ* 110:857–880
- Krugman P (1979) Increasing returns, monopolistic competition and international trade. *J Int Econ* 9(4):469–479
- Krugman P (1980) Scale economies, product differentiation, and the pattern of trade. *Am Econ Rev* 70(5):950–959
- Krugman P (1991) Increasing returns and economic geography. *J Polit Econ* 99:483–499
- Melitz MJ (2003) The impact of trade on intra-industry reallocation and aggregate industry productivity. *Econometrica* 71(6):1695–1725
- Ottaviano GIP (2003) Regional policy in the global economy: insights from the new economic geography. *Reg Stud* 37(6–7):665–673
- Ottaviano GIP, Tabuchi T, Thisse J-F (2002) Agglomeration and trade revisited. *Int Econ Rev* 43:409–435
- Puga D (1999) The rise and fall of regional inequalities. *Euro Econ Rev* 43(2):303–334
- Redding SJ (2010) The empirics of new economic geography. *J Reg Sci* 50(1):297–311
- Redding SJ, Sturm DM (2008) The costs of remoteness: evidence from German division and reunification. *Am Econ Rev* 98(5):1766–97
- Venables A (1996) Equilibrium locations of vertically linked industries. *Int Econ Rev* 37:341–359
- World Bank (2008) *World development report 2009*. World Bank, Washington

Further Reading

- Baldwin R, Forslid R, Martin PH, Ottaviano GIP, Robert-Nicoud F (2003) Economic geography and public policy. Princeton University Press, Princeton
- Combes P-P, Mayer T, Thisse J-F (2008b) Economic geography. Princeton University Press, Princeton
- Head K, Mayer TH (2004) The empirics of agglomeration and trade. In: Henderson V, Thisse JF (eds) Handbook of regional and urban economics, vol IV. North Holland, Amsterdam, pp 2609–2665
- Neary JP (2001) Of hype and hyperbolas: introducing the new economic geography. *J Econ Lit* 39(2):536–561

Evolutionary Economic Geography and Relational Geography

31

Harald Bathelt and Peng-Fei Li

Contents

31.1	Introduction	592
31.2	Segmented Cluster Paradigms	593
31.3	Network Relations and the Knowledge-Based Conception of Clusters	595
31.4	Regional Path Dependence and Cluster Life Cycles	596
31.5	Toward an Integrated Relational-Evolutionary Model of Cluster Dynamics	599
31.6	Conclusions	604
	References	606

Abstract

In the past decade, economic geography has encountered increasing interest and debates about evolutionary and relational thinking in regional development. Rather than comparing the two approaches, this chapter investigates how they can complement one another and be applied to specific research fields in economic geography. A comparison would be difficult because the approaches address different levels of the research process and are in a relatively early stage of their development. To demonstrate the potential of combining the two approaches, this chapter aims to conceptualize cluster dynamics in an integrated relational-evolutionary perspective. In recent years, research on clusters has experienced a paradigmatic shift from understanding their network structure to

H. Bathelt (✉)

Department of Political Science and Department of Geography & Program in Planning, University of Toronto, Toronto, ON, Canada

e-mail: harald.bathelt@utoronto.ca

P.-F. Li

Department of Urban & Regional Economy and Institute of China Innovation, East China Normal University, Shanghai, People's Republic of China

e-mail: pfli@re.ecnu.edu.cn

analyzing dynamic changes. Within this context, inspired by relational and evolutionary thinking, a comprehensive tripolar analytical framework of cluster evolution is developed that combines the three concepts of context, network, and action, allowing each to evolve in interaction with the others. Through this, the chapter argues that, rather than viewing relational and evolutionary accounts as competitive approaches to economic geography, they can, in an integrated form, become fundamental guides to economic geography research.

31.1 Introduction

After vivid conceptual debate in economic geography in the 2000s, two approaches have received substantial attention in the academic community that will be discussed in this chapter: that is, relational and evolutionary perspectives. While some scholars compare both perspectives as competing conceptualizations (e.g., Hassink and Klaerding 2009), we believe that such a comparison is not easily possible. There are two reasons for this: First, relational economic geography is a broad term which encompasses a number of approaches that relate to different research traditions stretching out from critical realist and poststructuralist to actor-network theorizations – which makes it difficult to critique this work as a homogenous body of research – while evolutionary economic geography has more narrowly developed out of evolutionary economics. Second, relational perspectives address meta-theoretical aspects of how to position analyses in economic geography, which questions to ask and how to conceptualize specific problems. In contrast, evolutionary approaches are situated at the concept level and often involve a specific quantitative methodology to analyze a problem. Both approaches are also in a relatively early stage of their development.

Relational perspectives were designed as a multidisciplinary alternative to narrow regional science approaches which were primarily based on conventional neoclassical economics. Such work aimed to explain economic landscapes by introducing spatial variables into economic models. Although conventional analyses did not always strictly follow this line of thinking, much of the work was characterized by a meso-/macro-perspective, the treatment of spatial entities as if they were actors (while neglecting the real actors, i.e., individuals, firms, and other organizations), a neglect of wider social relations, and a lack of process analysis. Relational approaches instead view economic action as social practice, conduct a microlevel reasoning, investigate the way how institutions stabilize economic relations, explore social and economic processes, and analyze the effects of global production and the connection between local and global scales (e.g., Boggs and Rantisi 2003).

In the conceptualization of Bathelt and Glückler (2011), which provides the reference point for the arguments presented here, relational economic geography is a meta-conceptualization for formulating research questions and conducting research in economic geography. This conceptualization provides a bottom-up logic of how economic action unfolds in a spatial perspective and leads to wider spatial patterns that can differ from place to place. This includes a structural and an evolutionary component: The structural component refers to the role of *context*. Accordingly, economic

agents are situated in structures of social relations from which they cannot easily separate (Granovetter 1985). Firms in clusters and global value chains are, for instance, embedded in networks of knowledge flows and supplier-producer-user relations which are key when making decisions about product changes. The evolutionary component refers to the fact that economic action is *path dependent*. Past decisions and traditions of social relations provide preconditions for today's actions and thus impact contemporary decision-making. At the same time, the relational approach rejects the assumption that such patterns can be extrapolated to the future. Economic action is seen as fundamentally open-ended and *contingent*, since agents are free to deviate from preexisting structures and development paths. From this, it is suggested that the complex underlying structures of organization, interaction, innovation, and evolution in a spatial perspective are at the core of enquiries in economic geography.

Similar to relational perspectives, evolutionary approaches are based on a critique of conventional research that is mostly static. In contrast, evolutionary approaches to economic geography aim to analyze dynamic changes in economic landscapes, often based on conceptualizations from evolutionary economics (Martin and Sunley 2006). Focusing at the regional (or national) level, much of the work analyzes the effects of processes of selection, mutation, variation, and chance on the development of firm populations. Within this context, recent work investigates processes of establishing regional variety and selecting alternatives from this variety. The idea behind this is that “related variety” between local/regional industry sectors enables spillover processes, supports innovation, and produces regional advantage (Frenken et al. 2007). Over time, selection processes lead to specific regional development paths. While research on the establishment of new trajectories is still at an early stage, an older stream of the literature analyzes path-dependent regional development and potential lock-in processes (Grabher 1993).

Although offering new insights for studies in economic geography, both perspectives have shortcomings: Much of the empirical work using a relational framework aims at understanding why specific economic networks exist, what the nature of social relations is, and why this differs from place to place, while neglecting the dynamics of such structures. Vice versa, evolutionary approaches focus on regional economic dynamics and the identification of trajectories at a meso-/macro-level, while neglecting the underlying structures of socioeconomic relations. In fact, although evolutionary approaches are often based on a firm perspective, the actual analysis addresses aggregates, such as regional structures and developments, and derives general statements about, for instance, the persistence of regional distributions. These differences are illustrated further in the next section which directs attention to industrial clusters as the unit of analysis.

31.2 Segmented Cluster Paradigms

Arguably, empirical and conceptual analyses of industrial agglomerations and clusters have been at the core of much of the work in economic geography over the past three decades. Within this context, relational and evolutionary approaches

have developed in two successive stages of the discussions of industrial clusters: Initially, academic interest was attracted by the robust growth of certain industrial districts, clusters, or regional innovation systems. (This early stage of cluster research is, in fact, only partially relational since much of this work lacks dynamic components. This may explain why some scholars interpret relational approaches as static (Martin and Sunley 2006; Hassink and Klaerding 2009).) A consensus about the structure of these competitive regions was that they combine economic activities and culture at the local level through untraded (aside from traded) linkages, echoing with the social-embeddedness argument in economic sociology (Granovetter 1985). It was argued that networks of local agents, which are often associated with mutual trust, provide a third way of governing economic relations beyond the dual structure of market and hierarchy. They generate regional prototypes of tacit knowledge where new knowledge is constantly being created and successfully shared (Malmberg and Maskell 2002). In more recent research, due to changes in regional configurations, such as the Third Italy (Hadjimichalis 2006) and Silicon Valley (Saxenian 2006), a transition has taken place from a static to a more evolutionary view of clusters. This has given rise to a new evolutionary approach on clusters that focuses on dynamic changes, drawing inspirations from concepts such as path-dependency, lock-in, and industry life cycles (Frenken et al. 2007).

Until now, the two dimensions - network and evolution - have remained relatively unconnected in the literature on industrial clusters. In both perspectives, broader levels of change in social networks and local culture and their impact on cluster evolution have rarely been discussed. This chapter argues that a close linkage between network dynamics and cluster evolution needs to be established to develop a coherent conceptualization of clusters. Without changes in networks and conventions, regional renaissance would barely be a “flash in the pan,” induced by temporary increases in demand. Signs of recovery would not lead to a succession of the evolutionary path or life cycle of a cluster. Without an evolutionary perspective, regional success would be determined by the existing local manufacturing culture (Gertler 2004), which would also provide a partial understanding.

To bridge the gap between narrow relational and evolutionary perspectives in cluster research, this chapter formulates a tripolar framework for the analysis of cluster dynamics through contextualized theoretical construction (Li et al. 2012). The tripolar framework builds on the pillars of context, network, and action, integrating them in an organic way at the local/regional level. This is not an attempt to establish a global model of cluster dynamics. Rather, by contextualizing social networks, we emphasize the possibility of network dynamics and, hence, varied effects of networks on local agency over time. As such, contextualized networks help explain and understand deeper transformations inside clusters. Furthermore, by placing networks in dynamic context-action configurations, we indicate how new cluster paths can be created through structuration processes that are initiated by local agents (Giddens 1984).

Following this agenda, this chapter is structured as follows: The next section discusses relational cluster conceptions that focus on the network paradigm, drawing particularly from the knowledge-based buzz-and-pipeline model. Then, we

present an overview of evolutionary cluster conceptions. From a critique of both types of approaches, we develop a reconceptualization of cluster dynamics in an integrated relational-evolutionary way, before presenting concluding remarks.

31.3 Network Relations and the Knowledge-Based Conception of Clusters

Traditionally, work on industrial agglomerations or regional industry clusters has emphasized the role of cost advantages, especially low transportation and transaction costs and close material linkages within such settings. Krugman (1991), for instance, stressed the importance of cost incentives for suppliers to locate close to an existing industrial agglomeration and advantages of agglomeration from a labor market perspective. As contributions by Storper and Salais (1997), Malmberg and Maskell (2002), and others have emphasized, however, it is necessary to go beyond cost factors to more fully understand the processes underlying regional specialization and concentration. In drawing on “localized capabilities” and “untraded interdependencies,” broader conceptualizations of regional clusters acknowledge the importance of socio-institutional settings, interfirm knowledge flows and interactive learning in regional innovation and growth.

From this understanding, a research tradition has developed that stresses the importance of network linkages and producer-user relations in clusters. Focusing on local interfirm linkages, Malmberg and Maskell (2002) emphasize the vertical and horizontal dimensions and relationships in clusters. While the former relationships refer to firms that are linked through input–output linkages and value-chain-based relations, the latter relate to firms that produce similar products and compete against one another. They learn by monitoring and comparing themselves with other firms.

Although Malmberg and Maskell (2002) point out that, in order to establish a theory of clusters, it would be necessary to understand which factors support the continued growth of clusters and how they are reproduced, much of the existing work has not developed a dynamic or evolutionary perspective. This is also reflected in the buzz-and-pipeline model of clusters (Bathelt et al. 2004), which suggests that the growth of a cluster depends on systematic linkages between its internal networks, conceptualized as “local buzz,” and its external knowledge and market environment, referred to as “global pipelines.” Within the cluster, specific information about technologies, markets, and strategies is exchanged in a variety of ways in planned and unplanned meetings. Based on a shared institutional background, firms learn how to interpret local buzz and make good use of it. Participation in this buzz does not require specific investments, since the firms are surrounded by a tight web of opinions, recommendations, judgments, and interpretations (Storper and Venables 2004).

While local buzz supports internal coherence, a cluster’s competitive success and growth strongly depends on its external linkages (Owen-Smith and Powell 2004). Since access to global or trans-local markets and knowledge is not free, considerable search efforts have to be undertaken to find the right

partners – a process that entails high investments and uncertainties. External relationships also require building trust, which is a timely and costly process. The buzz-and-pipeline model suggests that the local information and knowledge ecology is of only limited effect in the absence of trans-local connections (Bathelt et al. 2004). The more strongly the actors in a cluster are involved in establishing and maintaining external partnerships, the more information about new markets and technologies is pumped into the cluster's networks (Fitjar and Rodríguez-Pose 2011). Without this influx of external knowledge, there is a danger that firms miss out new opportunities or pin their hopes on the wrong technologies. Vice versa, without local buzz, the cluster's external pipelines are also of little use. Local buzz enables firms to rapidly filter out from the mass of external information those elements that are particularly important for the development of technologies (Bathelt and Glückler 2011).

Although related cluster approaches often draw on dynamic concepts, such as growth and reproducibility, they are mostly static in character. Such approaches focus on network aspects and do not conceptualize the genesis and evolution of clusters (Maskell and Malmberg 2007).

31.4 Regional Path Dependence and Cluster Life Cycles

The growing interest in cluster dynamics originates from the failure of conventional static models in explaining local crises and structural changes. A dilemma of such research is that localized benefits are expected to happen once clusters exist. The question of how cluster structures emerge in the first place is neglected in this work or viewed as an “individualistic” process. Since the factors that support a cluster's genesis may differ from those that support its ongoing growth (Bresnahan et al. 2001), a systematic conceptualization of clusters requires a dynamic component.

One strand of the literature on cluster dynamics focuses on the concepts of path dependence and lock-in related to evolutionary theories. A conceptual challenge when applying metaphors from evolutionary economics or evolutionary biology to economic geography is, of course, to justify the transferability of path-dependence explanations – originating from microlevel analysis of organizational behavior – to the aggregate local/regional level. A natural way of justifying the use of evolutionary ideas at the local level is to demonstrate that geography matters in the realization of path-dependent processes. Such processes – be it related to technological lock-in, externalities, or institutional inertia – do not occur in a spaceless world. The idea that a firm's interactive learning processes, strategic choices, and organizational routines are shaped by the local cultural and institutional environment has been repeatedly pointed out in the network tradition of cluster research. In this view, path dependence is associated with a place-dependent evolutionary process (Martin and Sunley 2006). Various empirical studies add to this argument by illustrating that regional path dependence can persist over a very long time period (Grabher 1993; Saxenian 2006).

Further theoretical exploration of evolutionary thinking in economic geography goes beyond preliminary claims of regional stability. History matters but does not determine future trajectories of clusters (Bathelt and Glückler 2011). Related to this, it appears that path dependence overaccentuates the continuity and stability of regional developments while discontinuities and structural crises, which are equally if not more important, are rarely conceptualized. To conceptualize structural change and path dependence in a consistent manner requires a different interpretation of clusters.

The traditional path-dependence model treats clusters as homogenous entities that form the unit of analysis. Even though social relations of firms are acknowledged in the localization process of a new path, the central focus is the overall intensity of local networks rather than their internal structure, let alone changes in the network structure of social relations. Since the position in a network impacts the kind of knowledge an agent can receive, a diversified set of agents is likely associated with more diversified structures of local knowledge flows and networks. Therefore, new path creation of clusters – an outcome driven by the interaction between agents – is more likely to occur in regions with varied structures (Frenken et al. 2007; Boschma and Iammarino 2009). The focus on the heterogeneous and diverse nature of regions revitalizes evolutionary thinking in economic geography. In a transformational context, therefore, re-bundling processes in a region without diversified networks may lead to the development of hollow instead of renewed clusters (Bathelt et al. 2004). By viewing regions as composite systems and drawing inspirations from evolutionary ideas in political science, Martin (2010) puts forward an alternative model of path dependence to highlight dynamic path processes. By inspecting interactions of agents in different network positions, Sydow et al. (2010) make an effort to combine conceptions of path dependence with Giddens' (1984) structuration theory, thus trying to disentangle the underlying agency processes of new path creation in clusters. The tripolar framework, developed in the next section, draws from a similar conceptualization.

Although contributions to new path creation complement interpretations of clusters as quasi-permanent structures, evolutionary perspectives are, thus far, still limited by a relatively narrow focus of analysis. In views of path dependence, singular interpretations dominate, whereby cluster dynamics are restricted both theoretically and empirically to industry, technology, or institutional structures. In contrast, aspects of the coevolution of interrelated economic, technological, institutional, and sociocultural arenas – “a key issue for further research” (Martin and Sunley 2006, 413) – are remarkably under-conceptualized. Singular views of evolutionary dynamics are also strong in regional analyses. In conventional studies, the evolution of clusters has primarily been explained at the local level, leading MacKinnon et al. (2010) to criticize evolutionary economic geography as neglecting social structure, labor relations, and capital accumulation at a broader macro-level. In globalized competition, especially in capital-intensive industries, influences at the national and international scale are indispensable to understand evolutionary processes. To go beyond a regional theorization of cluster evolution is also propelled by international technical communities that promote cooperation and competition between clusters (Saxenian 2006).

Other conceptualizations of cluster evolution draw on industry or product life-cycle theories (Klepper 1997). Industry life-cycle theories suggest that a dominant product design does not exist during the early stages of industrial development and that new technologies only flourish in selected areas. With increasing maturity, markets become more stable, knowledge gets codified, and dominant technologies emerge. Since communication of tacit knowledge and technological innovation are key features of innovative clusters, a natural corollary of life-cycle theories is that innovative clusters most likely develop in an early rather than a mature stage of industrial development (Iammarino and McCann 2006). The point of this argument is that the evolution of clusters corresponds with and follows from the technological paradigm of those industries that form their bases.

Cluster life cycles are often a regional version of industry life cycles. Strong emphasis on the role of technology in cluster dynamics in these approaches, however, adds an element of determinism to the explanation. When using technological paradigms to explain the rise or failure of industrial clusters, there is a danger of *a posteriori* reductionist reasoning. It is easy to explain technological changes of an industry when looking back, but difficult to foretell what will happen in the future. Accordingly, only after clusters succeed or fail can the rationality of a technology regime be fully understood. In practice, however, a change in technology is not an external factor that determines the cluster's evolution but the outcome of the interconnected nature of the firms' choices and actions along the dynamics of cluster development. Cluster-, industry-, and product life-cycle theories predict technological change in a deterministic manner rather than explaining the origins of technologies as resulting, for instance, from cluster innovation.

Other cycle conceptualizations are different in that they presume that cluster life cycles are existent rather than constructed. Related research pays attention to uncovering the characteristic forces at each stage of a cluster cycle. Accordingly, different forces have been identified, through which clusters move from one stage to another (Maskell and Malmberg 2007; Menzel and Fornahl 2009). Early discussions of this type of cluster cycles implied that clusters experience a unidirectional stage-to-stage development. To free clusters from such deterministic reasoning, Menzel and Fornahl (2009) add feedback loops allowing clusters to jump back to earlier stages during the sustainability or decline stages. However, such relaxation of stage rigidity only alleviates the mechanical characteristics of cluster cycles. It is problematic to assume that a single life cycle could cover the diverse trajectories of clusters in different real-world contexts. Even Martin and Sunley's (2011) recent attempt to conceptualize cluster dynamics as an adaptive-cycle model drawing on evolutionary ecology does not fully overcome the idea of a "natural" development trajectory. They suggest that cluster evolution proceeds through different stages that can lead to continuous cyclicity but also get stuck in stages of ongoing adaptation, stabilization, reorientation/renewal or decline and disappearance of existing clusters.

Although evolutionary and life-cycle conceptualizations of cluster dynamics draw from different origins, they share two aspects in their theoretical construction. First, in both discussions cluster dynamics are typically conceptualized from model

assumptions rather than derived from within their regional or national contexts. Forces driving cluster evolution are, in many models, situated at aggregate levels beyond the individual agent. Such a way of conceptualization risks overabstraction, potentially losing sight of interesting insights happening “on the ground.” Studies may thus dismiss the diversity of trajectories of cluster development.

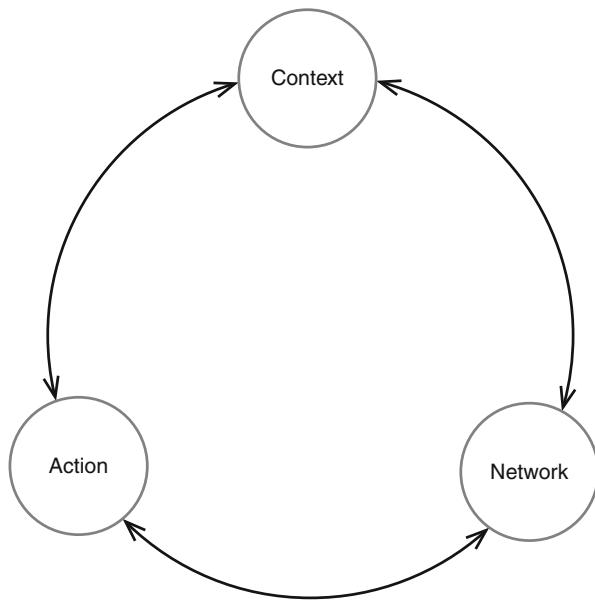
Second, there is an inclination to disregard changes in the underlying social structure. In cluster life cycles, different stages are mainly distinguished by observable indicators, such as firm size and number of employees, or by indicators that are less easily measurable, such as technology and diversity of local knowledge pools. In terms of the institutional dimension, it is institutional inertia rather than reforms of institutions that are captured by lock-in processes. Local business culture and social networks, which have been extensively discussed in the network paradigm of clusters, are deliberately excluded in these theoretical frameworks of cluster dynamics. It is suggested that path-dependent evolution leads to long-term stability in or irreversibility of spatial industry patterns (e.g., Boschma and Iammarino 2009). The problem of this view is that the seeming stability of aggregate patterns hides changes in the social structure and network relations underlying these meso-macro patterns.

As significant as evolutionary and life-cycle conceptualizations may be, their concentration on normative descriptions of cluster dynamics draws away attention from the analytical concerns of cluster theories. In a different approach, the theorization developed in the next section aims to frame the relationships of those forces enabling and shaping diverse trajectories of clusters. Based on observations of the economic agents’ behavior at the local level, as well as beyond, this conceptualization aims to extract key influences of cluster dynamics in the long run. Instead of asking how clusters will evolve, our analytical framework gives priority to the question of why clusters change. This does not presume the existence of a general theory that extracts the critical forces behind the dynamics of clusters.

31.5 Toward an Integrated Relational-Evolutionary Model of Cluster Dynamics

Any conceptualization of clusters presumes an interpretation of what a cluster is. In our view, a cluster is neither an organism, which can grow and decline per se, nor an entity, which can be described by a single rationality or technology. In the tripolar framework, a cluster is a group of agents and firms that are bound together geographically, technologically, and relationally. In this vein, trajectories of clusters are aggregate – planned as well as unanticipated – outcomes of the individual choices and actions of local agents, as well as the synergies that derive from them. Analytical frameworks of cluster dynamics need to be formulated in relation to the actions and motivations of local agents. From the contingent, relational, and accumulated characteristics of the local agent’s behaviors (Bathelt and Glückler 2011), three important pillars are identified as central analytical categories in the tripolar framework. These are context, network, and action, bound together in a reflexive manner that stimulates an evolutionary dynamic (Fig. 31.1).

Fig. 31.1 A tripolar analytical framework of cluster evolution (Source: Li et al. (2012, 133))



Context. Actions of local agents are contingent, which makes it hard to predict such actions. Contingency is directly related to the first pillar of our framework: the specific context in which actors are situated. By context, we mean the economic and institutional structures influencing local actors in the process of making and fulfilling decisions. This influence also includes the results of previous actions of other agents. The economic structure of clusters involves industry and market characteristics, technological patterns, intra-firm organization, and the dominant interfirm linkages inside and beyond the region. The institutional structure, in turn, refers to the local and nonlocal political regimes, routines, conventions, and value and belief systems. The economic and institutional settings, which are structured by the division of labor and the geographical distance between activities (Storper 2009), influence the local actors' knowledge base and their interaction. When applied to cluster evolution, the economic and institutional dimensions of context are often blurred since long-term interfirm connections can form powerful interest groups, stabilizing the local institutional context (Grabher 1993).

Context both constrains and enables action in clusters. From a psychological and pragmatic perspective, Storper (2009, 13) proposes an informational interpretation of context, the structural component of which "is defined by the division of labor in which the actor finds himself, which has a decisive influence on the information environment for the individual, hence his 'input' structure of cues and reference points." In this sense, context has an impact on the ways how actors find and apply information and knowledge, leading them to choose certain actions over others. The relationship between context and action is neither predetermined nor normative. A specific context does not determine what actors do but limits ways of coordinating actions in a given situation. In other words, there are different frameworks of action

in possible worlds of production, yet, in a certain context, some coordinated collective actions are more likely than others (Storper and Salais 1997). The effects of context on performance are not predetermined as they can be positive or negative (Storper 2009). On the negative side, the practical environment of actions restricts what kind of knowledge local actors may receive. On the positive side, context enables what agents in clusters can do by creating a bias toward certain kinds of knowledge. Therefore, for local agents, the question is not how to escape restricted contexts and/or enter more beneficial environments, but how to reflexively interpret practical situations and make appropriate adjustments. Context becomes an important influence once it has been internalized into the actors' motivations and behavior. In sum, the constitution of context reflects the duality of structure, both as a medium and an outcome of the agents' practices (Giddens 1984).

From a structural perspective, actions are structured by contexts. At a particular time for a specific local actor, context is a given constraint. Over a long time span, however, contexts are constructed by actions and are thus variable. Routines and conventions of doing business are formed based on foreseeable expectations about the mutual behavior of others as an outcome of recursive interaction. Ongoing interfirm relations are consolidated through the successes of series of transactions. Competitive patterns of industries are shaped by the choices and practices of actors in comparison with those of their competitors. Context is thus not a predestined background against which agents make choices and take actions; rather, it is constructed and sustained by ongoing practices of *all* agents. This means that the context, in which all agents are situated, usually cannot be controlled by single or exclusively local agents. At the cluster level, actors can modify their context in several ways, but there are also important components that are out of the hands of local agents. Although firms in clusters may engage in collective action to alter the supply conditions of a specific industry, they cannot easily change the demand of customers directly or influence national macroeconomic policies, legislative frameworks, and education systems.

Network. Network refers to the contextualized social relations of agents and firms within, but not limited to, the local production system. The wider structure of input–output linkages of firms also becomes part of the economic context of agents in clusters. In practice, social and economic relations are inseparable indicating that traded and untraded interdependencies (Storper and Salais 1997) are closely interwoven. As to trade linkages, the incompleteness of contracts leaves room for the development of trust (or distrust) between related partners in the negotiation and during the course of economic transactions. Mutual trusted relations become indispensable for traded interdependencies. Social networks in some places also originate out of economic rationales. Although not originally part of economic transactions, new personal relationships may be established over time through repeated economic transactions. In the end, however, it is the compellability and inspiration of personalities that trigger the formation of new social networks at a person-to-person level. Economic transactions offer opportunities for interaction and communication based on which some personal relationships develop, and not others. At the regional level, personal social networks may exist before the

formation of clusters. Such networks can become a key mechanism in the diffusion of market and technology information and develop into a cluster later on.

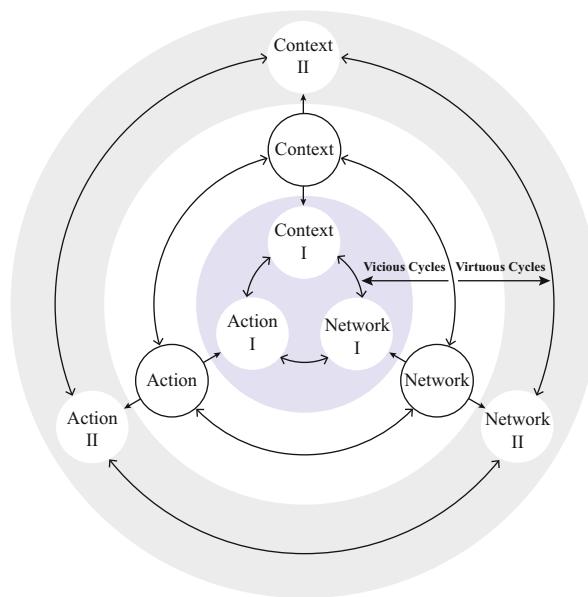
Changes in value and belief systems, advances in telecommunication technologies, and the intensification of interfirm competition can trigger a transformation of personal interaction toward a broad societal level beyond the region. In reflecting trust and ontological security systems between different societies, kinship relations can be viewed as providing a stable mode of organizing personal relations in the premodern societal context. These relations have been substituted by relationships of friendship or emotional intimacy in modern society. It is thus reasonable to assume that for clusters in developing or transitional economies, structures of social networks at the personal level will also change in the modernization process. Such a change of basic personal networks also impacts strategic actions within clusters, yet to maintain personal relations requires regular interaction and communication. Networks in this sense are “as much process as they are structure, being continually shaped and reshaped by the action of actors who are in turn constrained by the structural positions in which they find themselves” (Nohria 1992, 7). By viewing networks as dynamic connections within heterogeneous contexts that are shaped by actions, the context-network-action framework conceptualizes deeper changes in the socioeconomic structure of clusters.

Action. Even though context and network offer powerful insights into the behavior of local agents, action still needs to be treated as a separate pillar in our framework because experience from action develops in a cumulative fashion, and agents learn based on their absorptive capability (Cohen and Levinthal 1990). At both the individual and organizational levels, prior related knowledge helps and directs agents to use and assimilate new knowledge. The more specialized knowledge agents have previously acquired, the faster they can learn within their context or network. The role of absorptive capability of agents suggests that learning is a cumulative and path-dependent process with self-reinforcing characteristics. A conceptualization of cluster evolution without action would bear the risk of overemphasizing exogenous variables. Action refers to the individual level of decision-making that depends on specific personal and internal organizational structures, as opposed to the external context.

In our framework, context, network, and action are equally indispensable. Conceptualizing cluster dynamics without recognizing all three pillars provides only a partial understanding. Merely emphasizing the role of stable networks in local actions risks failing to understand diversifying patterns with transitional or developmental background. Conceptualizations, which limit themselves to emphasizing the importance of external contexts for actions, conversely neglect the role of human agency in regional practices (Scott 2006). Also, the theorization of actions that are withdrawn from the agents’ network and context would lead us to view clusters as organisms or groups of unrelated agents, neither of which would reflect real-world structures.

In sum, the tripolar framework offers a systematic way of studying and interpreting the evolution of clusters. At the regional level, it is the interaction of these pillars that explains the evolution of clusters, yet the framework does not

Fig. 31.2 Evolutionary dynamics in the tripolar cluster conception



produce ideal-type cluster visions since the dynamics of the three pillars can work in both vicious and virtuous ways:

- Vicious Cycles.* We refer to interrelationships between the pillars as being vicious if they produce lock-ins and result in regional decline, as illustrated by the contractive interactive movement of the three pillars in Fig. 31.2. In the literature, economic crises in industrial districts are often explained by changes in economic contexts, such as a sharp drop in demand or the appearance of new technologies. But a transformation of the external environment accounts for only one part in the overall stagnation of clusters. Weaknesses within the networks of a region can also be responsible for the rigidity of old industrial areas. Reasons for regional failure can be classified as different forms of lock-ins (Grabher 1993), which are consequences of interrelationships between the three pillars. First, decades of cooperation (action) in infrastructure projects and subsidy programs may stabilize intensive relations between people and firms (network) in an industry and corresponding policy field, thus strengthening a local conservative regime (context) that constrains further adjustment of local agents. The ossified institutional context may result in “political lock-in.” Second, long-term personal networks of local agents can result in similar reactions (action) to demand changes and technological opportunities. A homogeneous view of the world caused by intensive social networks in clusters may be the consequence of “cognitive lock-in.” Third, the stable demand for products may fixate the localized social division of labor and support a rigid economic context. The enduring fragmentation of activities among firms can result in shortcomings in the local agents’ learning processes and investment decisions regarding R&D. By exclusively concentrating on certain activities, the local agents’ accumulation of knowledge becomes biased and absorptive

capabilities with respect to new knowledge may become more restricted. In Grabher's (1993) classical typology, this rigidity of interfirm connections (economic context) generates "functional lock-in."

- b. *Virtuous Cycles.* In contrast to the above, virtuous interrelations between the three pillars can develop that have positive effects, as illustrated by the expansive interactive movement of the three pillars in Fig. 31.2. Agents with overlapping knowledge bases are, for instance, motivated to cooperate and communicate. Through the action and interaction of diversified agents, knowledge circulates in clusters, ideas collide, and innovation becomes more likely. In turning innovative ideas into business successes, the agents' relationships (network) that have been established in previous interaction are reinforced. Some commercialization of innovation may fail but there are also successes, which may reorder the existing industrial structure (economic context) and change the existing cluster path. Successful cooperation of agents not only results in economic returns to innovation but may also establish new interpretations of the context within which the agents are situated while producing important knowledge about their strengths and weaknesses. Agents with enhanced reflexive capability with respect to their context are more likely to act in anticipation of, rather than react to, future changes.

At the regional level, clusters with pre-active agents across different networks are characterized by high adaptability. This can lead to dynamic processes of path creation. Along with dynamic interactions of agents, mutual expectations regarding the coordination of actions may turn into unconscious routines and norms (institutional context), which become new components of the overall intangible regional assets. In the long run, these regional assets – both social networks and routines of doing business – are thus constructed, sustained, and altered through social reproduction. In Giddens' (1984) sense, the interaction of context, network, and action in virtuous cycles actively drives a regionalized structuration process.

31.6 Conclusions

This chapter started by discussing recent relational and evolutionary perspectives in economic geography, arguing that it is useful to integrate both approaches to combine their strengths rather than discussing them against one another. Applying these approaches to the study of regional industry clusters, it is suggested that both have shortcomings if used in isolation: While relational conceptualizations that focus on the social relations and structural dimensions of clusters tend to neglect aspects of cluster dynamics, evolutionary approaches do not sufficiently understand the underlying structure of social relations in clusters. To overcome this, we suggest a *tripolar framework of cluster evolution* that presents a combined relational-evolutionary perspective. Some elements of this framework are also reflected in the adaptive-cluster model described by Martin and Sunley (2011) – albeit at the

expense of assuming a predefined natural cycle. We believe that the tripolar approach provides important insights about network dynamics and cluster evolution in a spatial perspective:

First, the concept of *network* is relational in nature and should be interpreted in a contingent way (Bathelt and Glückler 2011). Research on networks in clusters has focused on the intensity of existing linkages, generally assuming that such ties are responsible for regional success or failure. Be the ties strong or weak, a relational-evolutionary perspective is skeptical of whether such a static interpretation of network can account for multifaceted regional developments. In the tripolar framework, network is only one pillar of the entire system and changes over time in interaction with the other pillars of context and action. One has to consider the dynamics of the *whole* framework to be able to properly evaluate the impact of strong or weak ties on cluster evolution. A specific network structure, for example, that supports a cluster's growth in one instance may turn out to be detrimental to regional competitiveness in a different setting.

Second, local traditions of action and interaction need to be evaluated in the specific *context* in which they matter for cluster evolution. Contextualized interpretations provide a perspective to understand why history matters in a nondeterministic way which is a thorny issue in evolutionary economic geography. With new political-economic contexts, for instance, new practices of interaction among individuals and organizations can form and become new elements of local structures and traditions. But not all practices of interaction develop into key elements of "regional assets." The degree to which a regional path can be established by local action depends on the specific context within which the local agents are situated (Storper 2009).

Third, the evolution of clusters is shaped by the aggregated *action* of local agents, as well as the unintended consequences of this action. Since some contexts are out of the hands of local agents, action may have unintended effects that shape future settings and affect individual and collective action in the next round. Consequently, the integrated relational-evolutionary framework rejects a normative model of cluster evolution or cluster life cycles, especially since there are also unexpected strategic actions that may, in the end, significantly alter the trajectory of clusters.

In sum, the tripolar framework conceptualizes cluster evolution through systematic interrelationships and ongoing feedbacks between context, network, and action. Focusing on the interdependencies of these important pillars, the framework demonstrates the value added of combining relational network-focused and evolutionary approaches in cluster research. A relational component in the tripolar framework helps explain *why* clusters evolve, thus avoiding deterministic elements in previous cyclical and evolutionary approaches. Further, an evolutionary perspective serves to extend the interpretation of local relations from a traditional static to a dynamic level of analysis. As an illustration of relational-evolutionary theorization on a specific topic, the tripolar framework reveals the potential of combining these different approaches to deepen our understandings of turbulent regional worlds. Therefore, this chapter may be regarded as an invitation to an integrated relational-evolutionary theorizing in economic geography.

Acknowledgements This chapter, to which both authors have contributed equally, is based on a more extensive conceptual and empirical study (Li et al. 2012). We would like to thank Andres Rodríguez-Pose for his encouragement and Manfred Fischer and Peter Maskell for thoughtful comments.

References

- Bathelt H, Glückler J (2011) *The relational economy: geographies of knowing and learning*. Oxford University Press, Oxford
- Bathelt H, Malmberg A, Maskell P (2004) Clusters and knowledge: local buzz, global pipelines and the process of knowledge creation. *Prog Hum Geogr* 28(1):31–56
- Boggs JS, Rantisi NM (2003) The ‘relational’ turn in economic geography. *J Econ Geogr* 3(2):109–116
- Boschma R, Iammarino S (2009) Related variety, trade linkages and regional growth in Italy. *Econ Geogr* 85(3):289–311
- Bresnahan T, Gambardella A, Saxenian A (2001) ‘Old economy’ inputs for ‘new economy’ outcomes: cluster formation in the new Silicon Valleys. *Ind Corp Change* 10(4):835–860
- Cohen M, Levinthal DA (1990) Absorptive capacity: a new perspective on learning and innovation. *Adm Sci Q* 35(1):128–152
- Fitjar RD, Rodríguez-Pose A (2011) Innovating in the periphery: firms, values and innovation in Southwest Norway. *Eur Plann Stud* 19(4):555–574
- Frenken K, van Oort FG, Verburg T (2007) Related variety, unrelated variety and regional economic growth. *Reg Stud* 41(5):685–697
- Gertler M (2004) Manufacturing culture: the institutional geography of industrial practice. Oxford University Press, Oxford
- Giddens A (1984) *The constitution of society: outline of the theory of structuration*. Polity, Cambridge
- Grabher G (1993) The weakness of strong ties: the ‘lock-in’ of regional development in the Ruhr area. In: Grabher G (ed) *The embedded firm: on the socio-economics of industrial networks*. Routledge, London, pp 255–278
- Granovetter M (1985) Economic action and social structure: the problem of embeddedness. *Am J Sociol* 91(3):481–510
- Hadjimichalis C (2006) The end of third Italy as we knew it? *Antipode* 38(1):82–106
- Hassink R, Klaerding C (2009) Relational and evolutionary economic geography: competing or complementary paradigms? *Papers in evolutionary economic geography* # 09.11, Urban & Regional Research Centre, Utrecht University, Utrecht
- Iammarino S, McCann P (2006) The structure and evolution of industrial clusters: transactions, technology and knowledge spillovers. *Res Policy* 35(7):1018–1036
- Klepper S (1997) Industry life cycles. *Ind Corp Change* 6(1):145–181
- Krugman P (1991) *Geography and trade*. MIT Press, Cambridge, MA
- Li P-F, Bathelt H, Wang J (2012) Network dynamics and cluster evolution: changing trajectories of the aluminium extrusion industry in Dali, China. *J Econ Geogr* 12(1):127–155
- MacKinnon D, Cumbers A, Pike A, Birch K, McMaster R (2010) Evolution in economic geography: institutions, political economy, and adaptation. *Econ Geogr* 85(2):129–150
- Malmberg A, Maskell P (2002) The elusive concept of localization economies: towards a knowledge-based theory of spatial clustering. *Environ Plann A* 34(3):429–449
- Martin R, Sunley P (2006) Path dependence and regional economic evolution. *J Econ Geogr* 6(4):395–437
- Martin R (2010) Rethinking regional path dependence: beyond lock-in to evolution. *Econ Geogr* 86(1):1–27
- Martin R, Sunley P (2011) Conceptualizing cluster evolution: beyond the life cycle model? *Reg Stud* 45(10):1299–1318

- Maskell P, Malmberg A (2007) Myopia, knowledge development and cluster evolution. *J Econ Geogr* 7(5):603–618
- Menzel M-P, Fornahl D (2009) Cluster life cycles – dimensions and rationales of cluster evolution. *Ind Corp Change* 19(1):205–238
- Nohria N (1992) Introduction: is a network perspective a useful way of studying organizations. In: Nohria N, Eccles RG (eds) Networks and organizations: structure, form, and action. Harvard University Press, Cambridge, pp 1–22
- Owen-Smith J, Powell WW (2004) Knowledge networks as channels and conduits: the effects of spillovers in the Boston biotechnology community. *Organ Sci* 15(1):2–21
- Saxenian A (2006) The new argonauts: regional advantage in a global economy. Harvard University Press, Cambridge
- Scott AJ (2006) Origins and growth of the Hollywood motion-picture industry: the first three decades. In: Braunerhjelm P, Feldman M (eds) Cluster genesis: technology-based industrial development. Oxford University Press, Oxford, pp 17–38
- Storper M (2009) Regional context and global trade. *Econ Geogr* 85(1):1–21
- Storper M, Salais R (1997) Worlds of production: the action framework of the economy. Harvard University Press, Cambridge, MA
- Storper M, Venables AJ (2004) Buzz: face-to-face contact and the urban economy. *J Econ Geogr* 4(4):351–370
- Sydow J, Lerch F, Staber U (2010) Planning for path dependence? The case of a network in the Berlin-Brandenburg optics cluster. *Econ Geogr* 86(2):173–195

Path Dependence and the Spatial Economy: A Key Concept in Retrospect and Prospect

32

Ron Martin

Contents

32.1	Introduction	610
32.2	Path Dependence as Self-reinforcing Spatial Economic “Lock-in”: What Does It Mean and How Common Is It?	612
32.3	Rethinking Path Dependence: From “Lock-in” to Ongoing Path Evolution	618
32.4	Toward a “Developmental–Evolutionary” Model of Path Dependence in the Spatial Economy	620
32.5	Conclusion	626
	References	628

Abstract

The concept of path dependence has rapidly assumed the status of a “fundamental principle” in the new paradigm of evolutionary economic geography that has emerged over the past few years. This chapter reviews the interpretation and use of this concept within this new field. The dominant interpretation has been that of “lock-in,” by self-reinforcing mechanisms, of particular (equilibrium) patterns of industrial location and regional specialization. This model is somewhat restrictive, however, and does not capture the full repertoire of ongoing path-dependent evolutionary trajectories that can be observed in the economic landscape. To respond to this limitation, the chapter suggests a “developmental–evolutionary” model of path dependence that includes “lock-in” as a special case, but which is also more general in its application and relevance.

R. Martin

Department of Geography, University of Cambridge, Cambridge, UK
e-mail: rml1@cam.ac.uk

32.1 Introduction

In recent years, the concept of path dependence has assumed a key explanatory role in a wide spectrum of social sciences and is now part of the standard lexicon of any approach that has pretensions to being “evolutionary” in orientation. Any evolutionary perspective on the socioeconomic starts from an elementary but important fact, namely, that, in each period, a socioeconomic inherits the legacy of its own past. Once this is acknowledged, we are faced with the possibility that “history matters.” The notion of path dependence is intended to imbue this idea with some degree of conceptual and explanatory rigor and hence go beyond simple narratives that merely describe historical effects. The idea of path dependence in its modern form was first developed by economists Paul David and Brian Arthur in the 1980s and early 1990s (David 1985, 1986; Arthur 1989, 1994) to explain technology adoption processes and industry evolution. Arthur’s discussion of the concept is particularly interesting, since he made explicit reference to the importance of path dependence in shaping the location of industry (Arthur 1994), a point also emphasized around the same time by Paul Krugman, who argued that:

If there is one single area of economics in which path dependence is unmistakable, it is in economic geography – the location of production in space. The long shadow cast by history over location is apparent at all scales, from the smallest to the largest. (Paul Krugman 1991, p. 80)

Although an early use of path dependence ideas to explain regional development was Grabher’s (1993) study of the Ruhr in Germany, it has only been over the past decade or so that the notion has really been taken up by economic geographers. In particular, the concept has assumed central importance in the theoretical and empirical contributions to the new paradigm of evolutionary economic geography that has emerged over the past few years (see Boschma and Martin 2007, 2010; Martin 2010a) and has come to be regarded as a key “organizing concept” for understanding how the economic landscape evolves over time.

The reason for this take-up of the concept of path dependence in evolutionary economic geography is not hard to explain. As evinced by David and Arthur, path dependence refers to a particular type of process that leads to the asymptotic convergence of an economic form or structure to a stable, “locked-in” configuration that can only be changed or “de-locked” by some sort of external shock or disturbance (see also Castaldi and Dosi 2006). Likewise, geographers argue, industrial location patterns and regional economic specializations show a similar process of self-reinforcing “lock-in.” We know that regional industrial structures, local economic specialisms, urban locations, and geographical patterns of development do not suddenly spring up over night, but have their origins in the past, and are built up over time, in many cases spanning several decades. Neither, typically, do spatial economic structures and configurations change suddenly. It is clear that any one point in time, the spatial structure of an economy is very similar to, and highly influenced by, the structure in the immediate and even less immediate past. The economic landscape we observe at any point in time has been shaped by the

historical adjustment path taken to it: it reflects its past development. Put another way, the economic landscape evolves as a consequence of its own history. In this sense, as Krugman argues (in the quote above), it is possible to argue that history casts a “long shadow” over industrial location patterns.

Yet, intuitive and appealing though this invocation of path dependence as a process shaping the spatial economy might be, it is not a straightforward notion. Even a “lock-in” definition can be given different interpretations and representations, and “lock-in,” if present, can be considered to be a positive feature or a negative one (see Martin and Sunley 2006). Further, different authors have used different formal models to represent path dependence, and these imply somewhat different definitions of the process. More generally, in the last few years, the frequent definition of path dependence as “lock-in” has itself come under increasing scrutiny, especially in political science, historical sociology, and management and organization studies. In these fields, there has been a growing reaction against the original model of path dependence as articulated by David and Arthur. Critics of this conceptualization argue that “lock-in” implies stability, stasis, or no change, or at the very least a particular form of evolution, namely, one in which long periods of stability are separated by periodic phases of rapid and disruptive change, whereas in reality many social and political structures and product and process technologies within business organizations evolve more or less continuously, yet still display path dependence. Accordingly, these writers have put forward alternative or revised models of path dependence that they believe more faithfully capture the actual varieties and patterns of path-dependent development observed in socioeconomic and technological systems. These alternative models may not produce the very “strong” form of path dependence associated with “lock-in,” but rather depict path dependence as an ongoing or “unfolding” process of adaptation whereby purposive behavior by individual agents, drawing on the structures and outcomes inherited from the past, actively reshapes those structures and outcomes: path dependence and path adaptation become inextricably linked.

Economic geographers and regional analysts do not seem to have been fully appreciative of these debates, yet they have important implications for how the concept of path dependence can be used to understand how regional and local economies evolve (see Martin and Sunley 2006; Martin 2010a, 2012). And here, additional issues also intervene. Regional and local economies are complex, heterogeneous, and highly open systems, often encompassing several different industries and activities, or subsystems, and are unlike the singular technologies, institutions, or products that are so often the subject of path dependence analysis in other disciplines. This complexity begs the question of what it is in regional and local economies that is path dependent. In addition, and potentially of critical importance, is the question whether path dependence is simply a general process or dynamic that shapes geographical economic outcomes, or is itself a process that is shaped by geographical context: in other words, is path dependence to some extent place dependent (Martin and Sunley 2006)? Still further, how does the concept of path dependence relate to our existing theories of (uneven) regional development? One of the criticisms leveled at the expanding paradigm of

evolutionary economic geography is that its advocates seem more intent on constructing a separate perspective than on integrating their evolutionary ideas and concepts with existing approaches, some of which, it is claimed, already take history seriously in one way or another (Mackinnon et al. 2009; Coe 2011; Oosterlynck 2012). This issue would seem to apply a fortiori to path dependence. Should the aim be to construct a distinct “path dependence theory” of regional growth and development, and what would such a theory look like? Or should the objective be to explore both the implications of the concept for existing regional theories, and what those theories imply for the idea of path dependence?

It is certainly not possible to take up all of these issues in detail in this short review of the “state of the art,” and what follows is necessarily somewhat selective and partial in coverage. I begin by summarizing the “canonical” interpretation of path dependence as “self-reinforcing lock-in,” and how far and in what ways this model is applicable to the spatial economic landscape. I then move on to examine some of the alternative views of path dependence that have been emerging in certain social and historical sciences, and what these interpretations imply for how we might think about the idea of regional path dependence. Building on this discussion, I then suggest what might be called a “developmental–evolutionary” view of path dependence, an interpretation that, while capable of including the basic “lock-in” model as a special case, allows for a much wider repertoire of evolutionary outcomes.

32.2 Path Dependence as Self-reinforcing Spatial Economic “Lock-in”: What Does It Mean and How Common Is It?

At least four formal models (and associated interpretations) of path dependence can be identified from the economic literature (see Table 32.1). One attempt to characterize the process of path dependence mathematically is in terms of a dynamic system which can be reduced to a difference equation in some key dependent variable, say X , which possesses a unit root, which means that the value of X_t in any period t embodies “memory” of its previous values and is thus dependent on its entire prior adjustment path. Such a unit root system does not converge to any equilibrium value or state: instead, the “long-run outcome” is defined as such by virtue of the temporal distance from some initial starting state, that is, from X_0 . Such an interpretation begs the question, of course, of what the “carrier of history” is that imbues the system or characteristic X with path dependence. Secondly, some authors (e.g., Setterfield 1998, 2009) provide this mechanism in terms of Kaldorian-type recursive cumulative causation models of national and regional growth, in which technological and institutional dynamics also play a key role. Again, in these structural difference equation models, no long-run equilibrium solution or state is implied. Thus far, models of this sort have not figured in economic geographers’ studies of spatial or regional path dependence, even though they would seem to offer a potentially useful avenue to explore. Instead, economic geographers have tended to rely almost entirely on the original notions of path dependence developed by David and Arthur.

Table 32.1 Four formal models of path dependence

Formal model representation	Typical application
1. <i>Path dependence as an absorbing Markov chain process</i> , in which a system has a probability transition matrix with more than one absorbing states ($p_{ii} = 1$), so that the system converges on a final distribution across states that depends in the initial (starting) distribution	Model implied in David's writings on the "lock-in" of technologies or institutional standards to historically fixed and unchanging forms, which may or may not be the most (market) efficient
2. <i>Path dependence as a nonlinear Polya (urn) stochastic process</i> in which the probability of an outcome of a given type in a given period increases the probability of generating that same outcome in the next period (a "proportions to probability" mapping). Model converges to an equilibrium distribution that is dependent on the initial (random) distribution	Model used by Arthur to generate the progressive "lock-in" of locational distributions of industries or cities to long-run stable spatial patterns that become self-reproducing
3. <i>Path dependence as a unit root process</i> in which a dynamic system can be reduced to a difference equation in the key dependent variable, say X , which possesses a unit root, which means that the value of X in any period t embodies "memory" of its previous values and is thus dependent on its prior adjustment path	Model used by various authors to study short- and long-run macro-dynamic phenomena. Such unit root models do capture at least one key aspect of path dependence, namely, the propensity for transitory random events shocks to have permanent effects
4. <i>Path dependence as a recursive cumulative causation process</i> in which recursive feedbacks among the structural components or relationships of a system reinforce a given trajectory or pattern of development. No long-term equilibrium is implied	Model used to generate Kaldorian-type, export-driven models of national and regional growth, with recursive dynamics (see Setterfield 2009), for example, $X_t \rightarrow Y_t \rightarrow Z_{t+1} \rightarrow X_{t+1}$ and so on

Basic to David's and Arthur's interpretations is the argument that rather than assuming an economy converges to a unique (pre-given) equilibrium irrespective of where it starts from – the approach taken by conventional economic theory – the nature of the (long-run) equilibrium an economy reaches depends on the process of getting there, and this will depend on some happenstance event in the past which then becomes selectively and progressively "locked-in" by some form or other of "self-reinforcing" mechanism. Thus, instead of a single equilibrium, there are multiple possible equilibria and which one the economy ends up in will depend on contingent events in the past. Here is David on the issue:

Small events of a random character—especially those occurring early on the path—are likely to figure significantly in 'selecting' one or other among the set of stable equilibria, or 'attractors.' (David 2005, p. 151)

The elaboration of theories around the core concept of path dependent dynamics...encourages and enables economists to entertain the possibility that, in place of a unique equilibrium-seeking dynamic, they should envisage a process that is seeking an historically-contingent equilibrium. (David 2007, p. 2)

Both David and Arthur conceptualize path dependence in terms of the limiting distributions of non-ergodic stochastic processes (see Table 32.1). In David's

Table 32.2 Processes-generating path-dependent lock-in

David's model ("Network externalities")	Arthur's model ("Increasing returns effects")
1. <i>Technical interrelatedness</i> (the reinforcing effects of complementarity and compatibility among the different components of a technology and its use)	1. <i>Large initial fixed setup costs</i> (in effect the inertia of sunk costs)
2. <i>Economies of scale</i> (the benefits associated with the increasing use of a technology – such as a decline in user costs – as the technology gains in acceptance relative to other systems)	2. <i>Dynamic learning effects</i> (learning by doing or using and learning by interaction tend to entail positive feedbacks)
3. <i>The quasi-irreversibility of investments</i> (the difficulties of switching technology-specific capital and human skills to alternative uses)	3. <i>Coordination effects</i> (which confer advantages to going along with other economic agents taking similar actions) 4. <i>Self-reinforcing expectations</i> (when the increased prevalence of a product, technology, process, or practice enhances beliefs of further prevalence)

Based on Martin (2010)

accounts, path dependence is likened to a Markov chain process with one or more absorbing states. Which absorbing state (long-run limiting equilibrium distribution) such a system will end up in will depend on where it started – “history matters” – but once in that state, the system cannot escape: it becomes “locked” into that particular equilibrium outcome. In Arthur’s formalization, path dependence is represented by a nonlinear Polya urn process, which possesses a multiplicity of possible stable fixed (equilibrium) outcomes (structures) of which one will be dynamically “selected” and “locked-in”:

Often there is a multiplicity of patterns that are candidates for long-term self-reinforcement: the cumulation of small events early on ‘pushes’ the dynamics into the orbit of one of these and thus ‘selects’ the structure that the system eventually locks into. (Arthur 1994, p. 33)

Arthur uses this model to show how industrial location can be interpreted as a path-dependent process that progressively “locks into” a stable, fixed distribution (of shares) of firms across regions. It is assumed that the autocatalytic or self-reinforcing mechanisms that generate path dependence (see Table 32.2 for the mechanisms identified by David and Arthur) have a spatial dimension, in that firms choosing where to locate are attracted by the presence of other firms in a region. Arthur describes two such models. In the “spin-off” version, the path-dependent geographical distribution of industry occurs through a process of local firm “spin-offs” from parent firms: This type of birth mechanism is argued to have characterized the US electronics and car industries. In the “agglomeration economies” version, if one region by chance gets off to a good start, its attractiveness and the probability that it will be chosen will be enhanced, further firms may then choose this region, and it becomes yet more attractive because of the emergence of various agglomeration economies and externalities, and the concentration of firms there becomes self-reinforcing. If such agglomeration economies are unbounded, then the model predicts that all of the firms in the industry will eventually end up in

one region: Arthur suggests Silicon Valley as a possible example of this sort of “locational monopoly.” A not dissimilar similar idea of self-reinforcing “lock-in” of a spatial economic structure into one of a number of possible multiple equilibrium patterns is to be found in NEG models. However, since such models are basically comparative static in nature and spatial agglomeration as one possible equilibrium outcome occurs instantaneously (for given assumptions about transport costs, wage functions, etc.), the often-made claim that these models incorporate “history” and “path dependence” is questionable (see Martin 2010b; Garretsen and Martin 2010).

The point about both David’s and Arthur’s models is that according to chance (combined in some cases with necessity, such as the spatial distribution of raw materials or natural resources), a different path-dependent spatial outcome might have been obtained with some other region becoming dominant. Thus, with different initial conditions (chance, random, or contingent events) combined with self-reinforcing path dependence effects, a multiplicity of possible equilibrium spatial economic structures can result. Which particular equilibrium (long-run) spatial economic structure becomes “locked-in” is assumed to remain fixed until such time that it is “de-locked” by a disturbance of one kind or another. (Likewise, in NEG models, a shock – such as a reduction in transport costs or a policy intervention – can “de-lock” one equilibrium spatial distribution of economic activity and move the system to a different equilibrium pattern, what NEG theorists refer to as “locational hysteresis.”) Construed in these terms, then, the economic landscape evolves by the emergence and “lock-in” of historically contingent, long-run equilibrium locational patterns of industrial activity and specialization that are periodically disrupted and, eventually, replaced by the path-dependent development of new historically contingent equilibrium patterns: in essence, an evolution characterized primarily by “punctuated equilibria” (David 2007, p. 187, explicitly aligns his path dependence model with this view of economic evolution).

How far does this version of path dependence capture real-world patterns of regional economic development? To some extent, this depends on what it is we are looking at. Many industrial location patterns and local industrial specializations, once established, often do seem to be subject to, or to give rise to, “self-reinforcing” mechanisms and processes that “lock” those patterns in (Table 32.3). However, this conception also raises several questions (Martin 2010a). For one thing, the model predicts a progressive “lock-in” to a long-run equilibrium stable state – in the sense of a stable pattern of localization of an industry or pattern of local sectoral specialization. Indeed, David (2005, 2007) actually refers to “path-dependent equilibrium economics.” But just how long is the “long run”? The problem with using a formal stochastic model (such as an absorbing Markov process or a nonlinear Polya urn process) to define path dependence is that there is no correspondence between the (logical) “convergence to equilibrium time” in such models and the real history of actual real-world processes of economic activity and development. Nor do real-world economies necessarily ever reach equilibrium states, even of a “historically contingent” kind. The idea of an equilibrium, or of multiple equilibria, is an imposed assumption, not a proven fact. As Setterfield (2003) argues,

Table 32.3 Some possible sources of regional path dependence

Source	Features
1. Natural resource based	Regional development path shaped and constrained by dependence on a particular raw material or resource
2. Sunk costs of local productive, physical, and infrastructural assets	Durability and quasi-irreversibility of local specialized capital equipment, infrastructures, or built forms
3. Local external economies of industrial specialization	Marshallian-type dynamic externalities, and both traded and untraded interdependencies associated with specialized local industrial districts or clusters
4. Local technological lock-in	Development of distinctive local technological regime or innovation system through local collective learning, cognitive inertia, knowledge transfers, and imitative behavior
5. Localized “spin-off” firm birth process	Local parent firms are sources of “spin-off” firms in similar or related activity, possibly supplying the original parent firms, leading to the buildup of a local industrial specialization or cluster
6. Agglomeration economies	Generalized self-reinforcing economies associated with spatial concentration of activity, including product and labor market effects, thick networks of suppliers, services and information
7. Interregional linkages and interdependencies	Development path in one region shaped by and dependent on development paths in another region, for example, because of interindustry linkages (such as acting as a specialist supplier of inputs to, or dependent on inputs supplied by, an industry in another region)

Based on Martin and Sunley (2006)

the very process of an actually existing economy approaching a stable long-run equilibrium position or state is itself likely to stimulate individual behavior – “innovation,” as he calls it – by economic actors to change their activities and thus prevent that economy from becoming fully locked into that particular equilibrium state. This is the sentiment expressed in another way by Metcalfe et al. (2006), who argue that because knowledge is constantly changing, capitalism can never be in a state of long-run equilibrium, but is instead in a state of constant “restlessness” in which some sort of innovation and structural change occurs more or less continuously. To be sure, such innovation and structural change may at times be slow and incremental, or at other times may occur in fits and starts, but the idea of an economy being in any type of stable or fixed equilibrium, essentially a state of stasis, is incompatible with the nature of capitalism as a dynamic, evolving system, as a process of continual “creative destruction.”

In what sense, then, is it possible to talk of spatial or regional “lock-in”? As mentioned above, from one vantage point, the *locational patterns* of industry and specialization can be viewed as being characterized by a certain degree of “spatial equilibrium” or “lock-in.” However, the longer, in historical time, the “long run” is specified or permitted to be, the less the spatial structure of the economy is likely to remain in a stable, unchanging state. Moreover, even if the *spatial patterns* of

industries across regions and locations are stable (“locked-in”), this need not suggest that the particular industries and specialisms *within* individual regions and locations are necessarily in a state of stasis or stability, with no product, technological, or organizational change. However, this is precisely what Krugman (1991) ignores in his discussion of the path dependence and “persistent dominance” of the US Manufacturing Belt. While US manufacturing has long been concentrated in a relatively small area of the North East and East North Central regions, the nature of the manufacturing activity conducted there has evolved considerably over time. The geographies of production may shift and change relatively slowly, but the firms and industries in a region may be characterized by significant ongoing endogenous change. Competition from other producers in other regions, or from other producers within the same region, is a constant source of pressure on a given region’s firms to upgrade and modernize their products, to introduce new variants or ranges of products, and to improve their productive efficiency through innovation. If a region’s firms fail to upgrade and innovate sufficiently, they face losing competitiveness and market share and even going out of business. This process need not occur suddenly: economic landscapes are littered with industrial districts and clusters that have undergone slow and protracted decline. Some such districts and clusters may eventually disappear altogether. Others, however, may survive and even undergo renewed success, sometimes much smaller in size and serving specialist niche markets, and in other instances undergoing renewed expansion based on shifting into related or complementary specialisms (for a discussion of the different evolutionary trajectories that clusters may undergo, see Martin and Sunley 2011). A “punctuated equilibrium” model of path-dependent regional “lock-in” may not in fact fit many actual regional and local experiences.

Yet further, unlike a singular technology or institutional standard (of the sort favored by David and many others in their discussions of path dependence), a given industry in a region (even if it is the only industry) is typically composed of numerous firms, among which there is bound to be heterogeneity, of products, production methods, innovativeness, business strategy, and so on. This heterogeneity or *variety* – or “composition” effect – also suggests that the idea of a regional industry becoming “locked” into a stable equilibrium state, a state of stasis, is unlikely to be the norm. Very specific local circumstances are required for this sort of outcome to occur, such as a local industry based entirely on a local natural resource, or when the firms in a local industry are closely linked by a very high degree of technological interrelatedness – for example, a form of production involving a detailed horizontal interfirrm division of labor – such that a change in one firm would require a change among all or almost all other firms, which might prevent any one single firm from changing in isolation. Such examples do obviously occur. But in most cases, the “lock-in” of locational patterns of industrial specialization *across space* by no means implies or leads to the technological or product “lock-in” of the firms *in individual places*.

The technological and product bases of firms, and thus industries, can and do change and develop over time, and the trajectories along which such development occurs can and do display path dependence, in that the improvements, innovations, and adaptations that firms make to their products and technologies invariably build

upon and are shaped by their existing products and technologies, which in their turn evolved out of previous versions. Essentially, then, as Page (2006) points out, it is possible to distinguish between two main types of path dependence: path dependence in which long-run *equilibria* depend on history, and path dependence in which *outcomes* are history dependent. Equilibrium path dependence is where the long-run distribution over outcomes depends on past outcomes: it is all about the historically contingent selection and self-reinforcing convergence to one of a number of possible limiting distributions over outcomes and links directly to the idea of progressive “lock-in.” Outcome path dependence is where the outcome in a period depends on past outcomes. Equilibrium path dependence implies outcome dependence, since if the long-run equilibrium distribution over outcomes depends on the past, then so must the outcomes in individual time periods. But outcome dependence does not imply equilibrium dependence: history matters in that current outcomes can be related, to some degree or other, to previous outcomes, and thus different previous outcomes are likely to have led to different present outcomes – the process is path dependent – but the path of outcomes over time need not converge to any long-run equilibrium or stable outcome. What this opens up is the possibility of a wider interpretation or conceptualization of path dependence, in which “lock-in” is only one, and a particularly “strong,” possibility, and in which pathways of technological, industrial, and regional development themselves evolve and unfold over time, in an outcome-dependent manner. This suggests the need for a wider conception of path dependence.

32.3 Rethinking Path Dependence: From “Lock-in” to Ongoing Path Evolution

Over the past few years, an increasing number of political scientists, historical sociologists, and management scientists have begun to explore what form(s) such a wider conception of path dependence might take. The “lock-in” model of path dependence has been frequently adopted in these disciplines to describe how political systems, social institutions, management practices, and the like evolve over time. But a growing corpus of empirical work has indicated that the evolution of these systems, structures, and organizations may in fact be much more ongoing and incremental than the “canonical” path dependence model would suggest. Put another way, it is argued that the standard path dependence model overemphasizes stability at the expense of ongoing change, mutation, and adaptation.

In historical sociology and political science, three main mechanisms have been suggested that operate at the micro-level to impart ongoing change to path-dependent institutional evolution: “layering,” “conversion,” and “recombination.” In a *layering* process, an institution or other such system changes gradually by the addition and accretion of new rules, procedures, or structures to what already exists. Each new “layer” (rule, etc.) constitutes only a small change of the institution as a whole, but this process can be cumulative over time so that while path dependent, the institution also evolves, leading to the mutation or even transformation of the institution’s fundamental nature. Not only does the addition of a new rule or

procedure to an institution depend on there being existing network externalities for its success, but this addition changes those externalities incrementally – and sometimes more substantially – in the process. Continuous incremental institutional change is thus both path dependent *and* path evolving.

A second process by which ongoing path-dependent change may occur in political and sociological systems is “conversion,” that is, the reorientation of an institution or other such system in terms of form or function, or both. Conversion can occur in two ways. First, the addition of new “layers” (new rules, procedures, and so forth) is itself a source of institutional conversion or reorientation, since the addition of new rules or procedures typically arises from the need or desire to alter an institution to serve new functions, roles, or imperatives. And the addition of a new layer may arise from, lead to, or necessitate the removal of an old layer. The second source of conversion is when the existing structures and arrangements of an institution are reoriented to serve new purposes, in response to external pressures or developments, or as part of a learning process by which existing rules are improved. No new rules or procedures as such need be added; rather, existing rules and procedures are realigned or modified. In some cases, however, the conversion of an institution may be possible only by means of a layering process. Although the recent political science literature has proposed that layering and conversion processes are separate and distinct mechanisms of incremental path-dependent institutional change, in reality they frequently coexist and interact. Moreover, while these mechanisms can be argued as alternatives to explanations couched in terms of path dependence mechanisms, they are in fact consistent with such mechanisms – indeed, they depend on them for their adoption and success. But unlike the canonical model of path dependence, layering and conversion processes need not lead to “lock-in”: Rather, they may well prevent “lock-in” from occurring.

Thirdly, other writers have proposed what they call a “recombinant” path dependence model. The basic idea is that any particular existing social–political–economic structure is, in effect, a system of resources and properties that actors can recombine and redefine, in conjunction with new resources and properties, to produce a new structure. Such a recombination is a source of path dependence in that what resources exist shape to some degree what changes can be made. The degree and nature of the “structured variety” that characterizes a socioeconomic system may thus be of some importance, since it will condition the range of resources that can be recombined. This recombination of existing social and institutional resources can be incremental but even plays a role at times of radical change.

In the management sciences, too, interest has focused on deriving alternative conceptions of path dependence that escape the restrictions of the canonical “lock-in” model and allow for ongoing evolution of a path. A key argument in this strand of literature is that standard path dependence models say little about *agency*, about how economic and other actors *create, recreate, and alter* paths (Garud and Karnøe 2001; Garud et al. 2010). The complaint is that the standard perspective on path dependence is that of an “outsider’s ontology”: the emphasis is on unpredictable contingencies, external increasing returns effects, and self-reinforcing nonlinear dynamics, which determine the behavior of actors who, once locked-in, cannot

escape unless some exogenous shock occurs. It is as if (local) economic actors are subject to some “higher-order logic” or “master plan” we call path dependence, over which they have no control. In contrast, it is argued, there is a need for another perspective on path dependence that embraces an “insider’s ontology,” that is, one which recognizes and assigns central importance to the purposive actions and behavior of actors. Purposive action is not only often responsible for the initial creation of a new path – which in the standard path dependence model is typically regarded as a happenstance or random event, “outside the knowledge” of the observer, or as Arthur (1994, p. 17) puts it, “beyond the resolving power of his [sic] ‘model’ or abstraction of the situation” – but also for how that path develops over time. Actors mobilize and draw upon the past (previous outcomes and experiences) in order to shape and fulfill their aspirations for the future: they may wish to repeat the past (to continue a particular form of activity) or to improve or move on from the past (by changing activity and behavior in some way). As Garud et al. (2010, p. 769) put it, “different visions of the future will lead to the mobilisation of the past in different ways. And these images of the future and mobilizations of the past will galvanise specific actions in the present.” Rather than “lock-in,” these authors argue, there is ever the possibility of creative destruction, with agents proactively innovating in order to move their activity forward under the pressure of competition and new opportunities. At the very least, economic agents learn, and contrary to the assumption made in the standard model of path dependence that learning leads to progressive imitation and to widespread adoption (i.e., “lock-in”), the assumption made by Arthur, learning can equally lead to more or less continual evolution of a path.

Although some, including Garud and his coauthors, want to go as far as to argue that path creation and path dependence should be regarded as distinct, others (such as Sydow et al. 2009) view path dependence and path creation as complementary and argue that any process is driven by a mix of the two. This opens up the possibility of different forms and degrees of path dependence, according to the balance of this mix of processes. Even “locked-in” states depend on agents’ decisions and actions – in this case, to change nothing and continue as before; whereas in many instances, there will be at least some agents, or groups of agents, whose intentional actions (entering into business, withdrawing from business, undertaking innovation, upgrading or redeveloping products, and so on) have the net effect that an industrial or technological or local economic path will mutate over time. In other words, the heterogeneity of decisions and actions among heterogeneous actors is very likely to prevent “lock-in” from occurring and instead lead to the ongoing adaptation of a path.

32.4 Toward a “Developmental-Evolutionary” Model of Path Dependence in the Spatial Economy

Elsewhere, I have suggested that these explorations into alternative perspectives on path dependence are highly suggestive for how we should think about the idea of

path dependence as a model of spatial economic evolution (Martin 2010a). I am not arguing that the processes operating in local or regional economies are identical to those shaping the development and evolution of institutional forms, merely that there are analogous processes at work in the former that resemble those in the latter, and that these are worth exploring and elaborating (see Martin 2012a). In the first place, while strong increasing returns effects may make for a self-reinforcing movement toward the “lock-in” of a particular industrial path in a local or regional setting, other mechanisms analogous to “layering,” “conversion,” and “recombination” may make for mutation and adaptation of that path over time. New firms are created or added more or less continuously as a local industry grows and develops; they may be spin-offs from existing firms, entirely new ventures, or implants from outside the locality. At the same time, some existing firms fail or move out of the locality. The addition and subtraction of competing entities and the consequential change in the relative frequency of different entities in a system are key forces that generate variety, and variety is a fundamental principle of evolution. New firms in an industry are likely to employ more advanced techniques, offer competing and perhaps different variants of the industry’s product or products, have different productivity and innovation profiles, and so on. The balance between entry, exit, and survival of firms may vary, of course, as the industry develops, and will be driven by a selection process that is determined, in large part, by the relative competitiveness of the firms in their relevant external markets.

Like “layering,” the idea of “conversion” also has an immediate relevance in a local industrial context. Changes to the characteristics of existing entities of a system are a key evolutionary mechanism. In the economic geographic case, it would refer to the ongoing innovation by firms in the local industry – in terms of new products, techniques, business organization, and the like – in response to market opportunities, competitive pressures, knowledge spillovers, and similar stimuli. The entry of new firms that employ newer techniques, different variants of products, and so on, by adding new elements to the local economy (i.e., “layering”), may, in turn, exercise a demonstration effect or spillover effect on existing firms leading to the “conversion” (reorientation) of their activities. As in the case of institutional evolution, these local industry “layering” and “conversion” processes interact, and “conversion” may well entail “recombination”-type processes, whereby firms are able to draw upon some aspects of existing local economic resources, capabilities, and externalities (such as skilled labor, technology centers, and the like) to reconfigure and reorient their activities. To the extent that mechanisms of these sorts operate, the technological and product “portfolio” of the local industry as a whole can change more or less continuously over time. Furthermore, as these changes cumulate, then the network externalities that support and benefit the local firms in the industry will also change. The skills of the local labor force, the range of intermediaries, of suppliers, of local supporting institutions – in fact, the whole gamut of local network externalities – may slowly evolve as the industrial path evolves. And driving the scale and direction of this evolution, of course, are the aspirations, reactions, and decisions of local actors, in firms, institutions, and other organizations.

This perspective on path dependence also allows for processes of “branching” to occur, whereby new, related sectors of activity emerge from and develop alongside, and perhaps even eventually replace, existing activities. The nature of such branching, itself a mechanism for creating variety or diversity in the local economy, will be path dependent to the extent that the new activities draw on competencies, technologies, and knowledges transferred from existing firms and activities. There is growing evidence that local economic diversification, the emergence of new activities and specializations, is often shaped by the existing industrial structure of an area, both in terms of influencing the scope for such diversification and the form it takes. In this sense, some local economic structures may be more “enabling” environments for branching and diversification to occur than others, which might be “constraining” in this regard. The point is that not only may a given local industrial path evolve and adapt over time but that a local economy’s entire economic structure may evolve in a path-dependent manner.

Conceptually, this can be thought of as a “developmental–evolutionary” model of path dependence in which there is a recursive relationship between the sectoral, technological, and institutional structures of a local economy, on the one hand, and the processes that drive economic evolution as these operate within and upon the local economy, on the other (see [Fig. 32.1](#)). At any one moment in time, a local economy will consist of a particular population of firms and businesses in specific sectors of activity and specialization, characterized by a range of technologies and processes, employing workers with certain skills, and linked to and regulated by, to varying degrees, local institutional arrangements. This local industrial–technological–institutional ecology provides the setting, the context, within which various processes of economic evolution operate. These processes depend on the conditions and circumstances obtaining in the local economy itself but also on various external factors, such as competitive pressures, new market opportunities, linkages to extra-local firms, external technological developments, regulatory norms, economic policies, and the like. The range of such external factors and developments that are relevant to and the particular effect they have on a local economy will itself also depend, in part at least, on that locality’s existing economic and technological structures. In combination, these local structures and external factors will stimulate, influence the scope for, shape the direction of, and constrain the mechanisms that make for change versus continuity within the firms and industries in the locality (the processes akin to “layering,” “conversion,” and “recombination” referred to above, namely, the local entry and exit of firms, the pace and direction of innovation by local firms, and the emergence of new or related activities). As the local economy’s sectoral and technological structure changes in response to these evolutionary processes, this then alters the ecology within which those processes operate, which then produces further change (or continuity) in the local economy’s development path, and so on.

The recursive model set out in [Fig. 32.1](#) is obviously a highly simplified representation of what in reality is a highly complex set of processes that can operate at different historical speeds across different firms and sectors, and at different spatial scales. But the key idea behind this model of path dependence is

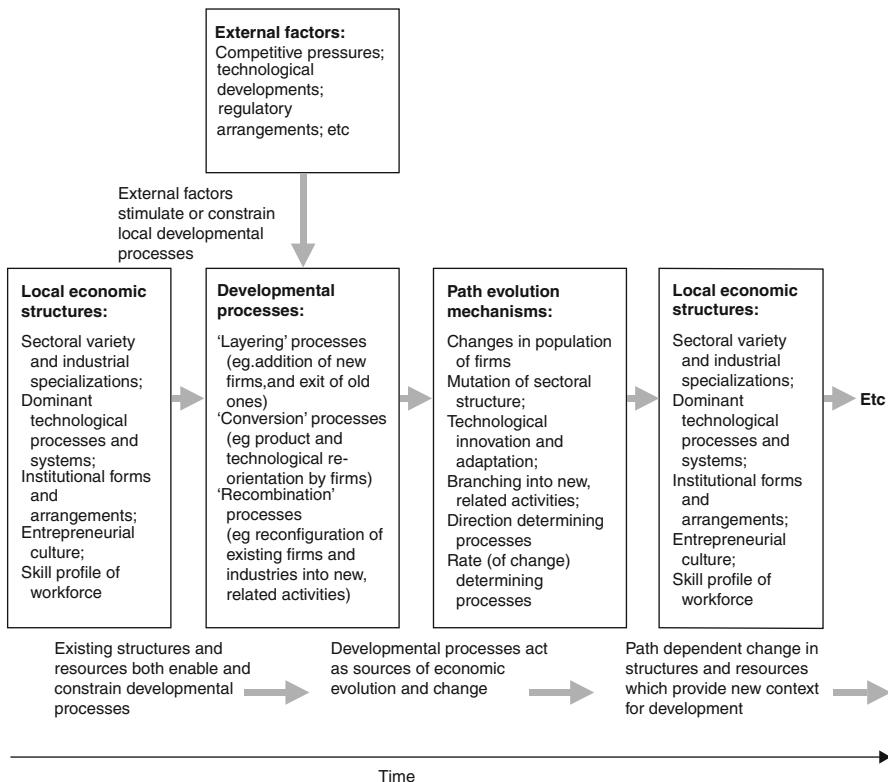


Fig. 32.1 A “developmental–evolutionary” model of local economic path dependence

that existing local economic and associated structures condition and influence the scope for and nature of developmental processes, which in turn shape the pace and direction of change in those structures, which in their turn feedback to condition and influence local economic developmental processes. Such a system is recursive and path dependent, but also potentially evolutionary. As such it offers a wider interpretation of path dependence than the “lock-in” model while encapsulating the latter as a special case (see Table 32.4).

Not only the scope for but also the pace at which local industrial paths evolve will obviously vary from one industry to another and from place to place. The slower the process, the more it is possible to describe the path or locational pattern as conforming to the standard “lock-in” model of path dependence: the research task in such cases is to determine why a local industry has failed to adapt and evolve. The faster the adaptation process, the less the notion of “lock-in” seems appropriate, the more the path will mutate in one way or another, and the more it is possible to describe the process as one of “developmental–evolutionary” path dependence. Allowing for purposive and intentional action by economic agents,

Table 32.4 The “lock-in” and “developmental–evolutionary” models of path dependence compared

	“Lock-in” model	“Developmental–evolutionary” model
1. <i>Initial conditions</i>	Inherited market and local economic and technological conditions assumed unimportant for creation of paths (“virgin market” assumption and “windows of locational opportunity” assumed open)	Inherited and constructed. Preexisting market and local conditions can enable or constrain new economic and technological possibilities. Previous and existing paths condition possibilities for new ones
2. <i>Contingencies</i>	Exogenous and manifest as unpredictable, nonpurposive, and somewhat random events	Emergent and serving as the embedded contexts for ongoing action by agents
3. <i>Self-reinforcing mechanisms</i>	Assumed key and given. “Systemic” mechanisms and processes that compel and constrain local agents’ decisions, which are largely beyond their control	May or may not be present locally. Not essential for path dependence and can be strategically manipulated by local (and extra-local) actors
4. <i>Lock-in</i>	Progressive convergence to a particular spatial distribution of an industry or local industrial structure or specialism, which is assumed to be an equilibrium state and unchanging until disrupted by an external shock	Lock-in (and equilibrium) is a special case. Most industrial location patterns and local industries undergo some form of ongoing change, adaptation, and evolution, both within industry (products, processes) and between industries (changing structural composition of local economy)
5. <i>Path plasticity</i>	While sometimes acknowledged (e.g., David 2001), lock-in to a stable outcome or path generally assumed to be the norm	Local industrial and technological paths can and do evolve incrementally. Actors constantly seeking improved and new product and process opportunities, and the cumulation of such behavior imparts mutation to an industry or specialism
6. <i>Path decline/destruction</i>	Assumed that some sort of (external) shock is necessary to “de-lock” a local industrial or technological path (state). Little discussion of gradual or long-run processes of decline of an industry or technology	Local industries and technologies display various types (and speeds) of relative and absolute decline. Causes of decline involve the complex interaction of exogenous and endogenous factors (e.g., failure or slowness of local firms to innovate and adapt in the context of changing market and competitive conditions)
7. <i>New path creation</i>	Not well theorized. What and where new industries and technologies emerge are contingent events	Constructed. New paths are typically the outcome of purposive and deviating behavior by agents, often influenced by or dependent on preexisting local conditions
8. <i>Model of economic evolution</i>	Punctuated equilibrium, whereby phases of stability (lock-in) are periodically disrupted by shocks (e.g., introduction of new technology)	Mutational and adaptive, allowing for incremental as well as discontinuous change

Note to table: In compiling Table 32.4, I have been influenced by Garud et al. (2010). However, their comparison is between the “lock-in” model of path dependence and what they regard as a distinctly different perspective on industrial and technological development which they call “path creation.” Since path dependence is an ongoing (re)creative process, driven by the activities and decisions of agents, these authors’ counterposition of path dependence and path creation is not perhaps that helpful. I prefer to follow Sydow et al. (2009) here, in seeing path dependence and path creation as inextricably interlinked and interactive, in an ongoing developmental–evolutionary process

for the normal ongoing processes of firm population dynamics, and for innovation, competition, and entrepreneurship sets a “developmental–evolutionary” model of path dependence apart from the standard conception. A “developmental–evolutionary” model is more admissive of the complex range of actual evolutionary paths that are found among industries, technologies, and local and regional economies.

Such a model also provides a richer perspective on the issue of new path creation. The standard model of path dependence does not have much to say about how, or where, new industrial–technological paths come into being, other than ascribing the emergence of new industries or technologies, and their locational geographies, to random, happenstance, or serendipitous events. Witt (2003) has questioned the validity, or at least the generality, of this “virgin market” idea, the assumption that the emergence of a new technology, product, or industry, and any competition with other emergent rivals, takes place without reference to and uninfluenced by inherited market conditions. Likewise, a “virgin landscape” assumption, the idea that where a new industry or technology emerges is unrelated to preexisting regional industrial and technological structures, can be challenged. There is a curious contradiction in the standard path dependence model, in that path dependence seems to matter only *once* a new industry or technology has emerged but plays no part in influencing *where* it emerges. In fact, there is growing evidence in economic geography that the inherited, preexisting industrial structure of a region or locality often does have an influence on whether a new industrial path emerges or develops there. A “developmental–evolutionary” model of path dependence, by giving recognition to developmental processes such as “layering,” “conversion,” and “recombination,” and how these are the outcomes of the purposive and intentional behaviors of local economic actors, admits of several possible mechanisms by which new local industrial pathways of economic development can emerge from preexisting ones. New paths can emerge from old. Local new path creation may thus itself form part of the developmental–evolutionary path dependence process: local preexisting capabilities, competencies, technologies, and knowledges can provide a resource base for local actors to deliberately venture into new or related fields.

Similarly, a “developmental–evolutionary” model offers a wider perspective on how local industrial and technological paths come to an end. In the standard “lock-in” model, it is assumed that a path is broken by some unpredictable external “shock.” Although such shocks can and do occur, and can certainly undermine or disrupt a local industry, and perhaps lead to its decline and demise, to attribute the decline of an industrial or technological path invariably to the impact of some unexpected or unpredictable exogenous shock is not especially enlightening. To be sure, a local industry is not a closed system and is subject to a variety of external pressures (and new opportunities). But such pressures and challenges are more or less constant features of modern economic life and not necessarily spasmodic, infrequent events. What matters, therefore, is the nature of the pressures that impinge on a local industry, and how the industry reacts to them, which, in turn, depends on the industry’s resilience and adaptability

(Hassink 2010; Martin 2012b). Furthermore, the decline of a local industrial path may arise endogenously, for example, because of the exhaustion of innovation by local firms, which then become uncompetitive and decline, so the industry shrinks. It may also occur if local firms switch to a different, perhaps related, sector of activity on a new path that is perceived as affording more profitable opportunities. Martin and Sunley (2006) suggested a number of possible mechanisms by which a local industrial path may be disrupted, or even destroyed, most of which revolve around the interaction of exogenous and endogenous forces. Yet, their analysis, like that of Castaldi and Dosi (2006) and, indeed, like many of those in economic geography, was founded on the assumption that the problem is one of identifying the mechanisms by which a “locked-in” stable state can be “de-locked.” But if “lock-in” never occurs, clearly a different conceptualization is needed as to how industrial development paths lose momentum, atrophy, and decline.

In several respects, then, a “developmental–evolutionary” model of path dependence differs significantly from the standard “lock-in” model. Whether it be in terms of the initial conditions that influence when and where a new industry or technology emerges, or the mechanisms that give rise to its path-dependent development across space, or the processes by which new industries replace old, or the type of local industrial evolution that is implied, the “developmental–evolutionary” model would seem more encompassing of the range of local economic evolutionary trajectories actually encountered. Furthermore, it incorporates the standard or strict “lock-in” model of path dependence as a special case. Thus, the notion of path dependence should be seen as a complex process that can take varying forms and produce varying rates of local industrial–technological change and evolution. In fact, the “developmental–evolutionary” model focuses attention precisely on why industrial paths vary in their evolvability (capacity to generate variety, of products, technologies, and indeed whole new industries) and adaptability and why these processes vary from one local economy to another.

32.5 Conclusion

According to Boschma and Frenken (2006, p. 280–281), “evolutionary theory deals with *path dependent* processes, in which previous events affect the probability of future events to occur” (emphasis in the original). The original formulations of path dependence in economics focused on a model that defined the process as the progressive “lock-in” – to a stable state – of particular technologies. This model has been taken up in various disciplines, including economic geography, where it has been used to explain the emergence and self-reinforcing spatial localization of industries. However, the “lock-in” model represents a very restricted model of spatial economic evolution. In recent years, this restricted definition or interpretation has been increasingly questioned. In evolutionary economics itself, there have

been dissenting voices. Thus, Witt (2003, p. 124), a leading evolutionary economist, has argued that the notion of “lock-in” is antithetical to industrial and technological evolution:

[S]ome doubts should be raised about the plausibility of both the theoretical underpinnings of, and the empirical evidence for, technological or industrial ‘lock-in’... sooner or later there will always be new rivals who threaten the market dominance of a technology or variant. The erosion of market dominance under competitive pressure by new technologies supports Schumpeter’s empirical generalisation that an incessant process of creative destruction characterises modern industrial capitalism (Witt 2003).

The same argument must surely apply to local and regional economies. The issue, then, is whether the notion of path dependence should be narrowly restricted to and reserved for situations of true “lock-in,” as recently argued by Vergne and Durand (2010), or whether the concept can be meaningfully widened to incorporate processes and systems – including local economies – that display ongoing developmental evolution.

Adherents of the narrow, “lock-in” view will no doubt see this idea of a “developmental–evolutionary” model of path dependence as too broad and lacking definitional precision and analytical formalism, perhaps even not as path dependence at all. The problem with this reaction is not only that the empirical applicability of the standard or canonical “lock-in” model would seem to be limited, since many industries and technologies – and most local and regional economies – exhibit varying degrees of ongoing development and adaptation, but that such ongoing development is itself characterized by some degree of path dependence and thus requires conceptualization: we need a framework for analyzing how and why local economies differ in the rate and direction of path-dependent adaptation and evolution. What precise form such a framework will take is a task for future research, but it will entail linking path dependence much more closely with other evolutionary concepts used in economic geography. The “developmental–evolutionary” model proposed above offers considerable scope in this direction. It also entails methodological deliberation. One of the possible attractions of the standard “lock-in” model of path dependence, especially perhaps to economists, is that it can be given a formal (i.e., mathematical and equilibrium) representation (as outlined in Table 32.1 above). But formalism can come at a cost: it can close off (“lock-out,” one is tempted to say) empirical patterns and outcomes that do not fit the prevailing model yet which, if subjected to appreciative theorizing, could well suggest other interpretations and generalizations that are more relevant. The “developmental–evolutionary” perspective on spatial economic path dependence suggested here is intended to encourage greater appreciative theorizing from concrete cases in order precisely to widen the applicability of the notion. In this sense, future work on path dependence in the spatial economy might well resemble the new generation of “history-friendly” models that are being pioneered in evolutionary economics, in which it is explicitly recognized that industries and technologies can take a variety of evolutionary paths and where appreciative theorizing from concrete case studies is used to construct models that take full account of that variety (see, e.g., Malerba 2010).

References

- Arthur WB (1989) Competing technologies, increasing returns, and 'lock-in' by historical events. *Econ J* 99(March):116–131
- Arthur WB (1994) Industry location patterns and the importance of history. In: Arthur WB (ed) Increasing returns and path dependence in the economy. Michigan University Press, Michigan, pp 49–68
- Boschma R, Frenken K (2006) Why is economic geography not an evolutionary science? *J Econ Geogr* 6(3):272–302
- Boschma R, Martin RL (2007) Constructing an evolutionary economic geography. *J Econ Geogr* 7(5):537–548, Special Issue: Evolutionary Economic Geography
- Boschma R, Martin RL (eds) (2010) Handbook of evolutionary economic geography. Edward Elgar, Cheltenham
- Castaldi C, Dosi G (2006) The grip of history and the scope for novelty: some results and open questions on path dependence in economic processes. In: Wimmer A, Kössler R (eds) Understanding change: models, methodologies and metaphors. Palgrave Macmillan, London, pp 99–128
- Coe N (2011) Geographies of production, 1: an evolutionary revolution? *Prog Hum Geogr* 35(1):81–91
- David PA (1985) Clio and the economics of QWERTY. *Am Econ Rev* 75(2):332–337
- David PA (1986) Understanding the economics of QWERTY: the necessity of history. In: Parket WN (ed) Economic history and the modern economics. Blackwell, Oxford, pp 30–49
- David PA (2005) Path dependence in economic processes: implications for policy analysis in dynamical systems contexts. In: Dopfer K (ed) The evolutionary foundations of economics. Cambridge University Press, Cambridge, pp 151–194
- David PA (2007) Path dependence: a foundational concept for historical social science. *Cliometrica* 1(2):91–114
- Garretsen H, Martin RL (2010) Rethinking (new) economic geography models: taking geography and history more seriously. *Spat Econ Anal* 5(2):127–160
- Garud R, Karnøe P (2001) Path creation as a process of mindful deviation. In: Garud R, Karnøe P (eds) Path dependence and creation. Lawrence Erlbaum, London, pp 1–38
- Garud R, Kumaraswamy A, Karnøe P (2010) Path dependence or path creation? *J Manage Stud* 47(4):760–774
- Grabher G (ed) (1993) The weakness of strong ties: the lock-in of regional development in the ruhr area. The embedded firm. Routledge, London, pp 255–277
- Hassink R (2010) Regional resilience: a promising concept to explain differences in regional economic adaptability? *Camb J Reg Econ Soc* 3(1):45–58
- Krugman P (1991) History and industry location: the case of the manufacturing belt. *Am Econ Rev* 81(2):80–83
- MacKinnon D, Cumbers A, Pike A, Birch K, McMaster R (2009) Evolution in economic geography: institutions, political economy, and adaptation. *Econ Geogr* 85(2):129–150
- Malerba F (2010) Industry evolution and history-friendly models. Plenary paper, international schumpeter society conference on innovation, organisation, sustainability and crisis, Aalborg, 21–24 June 2010. <http://www.schumpeter2010.dk/index.php/schumpeter/schumpeter2010/paper/viewFile/491/208>. Accessed June 2011
- Martin RL (2010a) Roepke lecture in economic geography – rethinking regional path dependence: beyond lock-in to evolution. *Econ Geogr* 86(1):1–27
- Martin RL (2010b) The 'new economic geography': credible models of the economic landscape? In: Lee R, Leyshon A, McDowell L, Sunley P (eds) The Sage companion to economic geography. Sage, London, pp 53–72
- Martin RL (2012a) (Re)Placing path dependence: a response to the debate. *Int J Urban Reg Res* 36(1):179–192

- Martin RL (2012b) Regional economic resilience, hysteresis and recessionary shocks. *J Econ Geogr* 12(1):1–32
- Martin RL, Sunley PJ (2006) Path dependence and regional economic evolution. *J Econ Geogr* 6(4):395–437
- Martin RL, Sunley PJ (2011) Conceptualising cluster evolution: beyond the life cycle model? *Reg Stud* 45(10):1295–1318, (with P.J. Sunley) Special Issue on Cluster Life Cycles, Eds. R. Boschma and D. Fornahl
- Metcalfe JS, Foster J, Ramlogan R (2006) Adaptive economic growth. *Camb J Econ* 30(1):7–32
- Oosterlynck S (2012) Path dependence: a political economy perspective. *Int J Urban Reg Res* 36(1):158–165
- Page S (2006) Path dependence. *Q J Polit Sci* 1:87–115
- Setterfield M (1998) Rapid growth and relative decline: modelling macroeconomic dynamics with hysteresis. Macmillan, London
- Setterfield M (2009) Path dependency, hysteresis and macrodynamics. In: Arestis P, Sawyer M (eds) Path dependency and macroeconomics. Palgrave Macmillan, London, pp 37–79
- Sydow J, Schreyogg G, Koch J (2009) Organisational path dependence: opening the black box. *Acad Manage Rev* 34(4):689–709
- Vergne J, Durand R (2010) The missing link between the theory and empirics of path dependence: conceptual clarification, testability issues and methodological implications. *J Manage Stud* 47(4):736–759
- Witt U (2003) The evolving economy. Edward Elgar, Cheltenham

Gilles Duranton

Contents

33.1	Introduction	632
33.2	Cities, Worker Productivity, and Wages	632
33.3	Firm Dynamics Within Cities	638
33.4	City Functionality, Urban Systems, and Policies	642
33.5	Conclusions	646
	References	647

Abstract

This chapter discusses the literature on agglomeration economies from the perspective of jobs and job dynamics. It provides a partial review of the empirical evidence on agglomeration externalities; the functionality of cities; the dynamic relationship between cities, jobs, and firms; and the linkages between cities. We provide the following conclusions. First, agglomeration effects are quantitatively important and pervasive. Second, the productive advantage of large cities is constantly eroded and needs to be sustained by new job creations and innovations. Third, this process of creative destruction in cities, which is fundamental for aggregate growth, is determined by the characteristics of urban systems and broader institutional features. We highlight important differences between developing countries and more advanced economies. A major challenge for developing countries is the transformation of their urban systems into drivers of economic growth.

G. Duranton

Department of Economics, University of Toronto, Toronto, ON, Canada

e-mail: duranton@wharton.upenn.edu

33.1 Introduction

This chapter reviews the literature on agglomeration economies from the perspective of jobs and labor markets.

In cities, jobs are more productive because of agglomeration effects. These take place through a variety of channels: resource sharing, quicker and better matching, and greater knowledge spillovers. [Section 33.2](#) provides a discussion of these issues. The bottom line is straightforward; cities have a positive effect on productivity and wages.

More productive urban jobs however do not come in a void. [Section 33.3](#) broadens the discussion to job creation and firm dynamics in cities. More productive jobs in cities need to be created. Innovation, entrepreneurial activity, and firm growth all play a crucial role in this respect. Adding to this, more productive jobs do not remain more productive forever. This productivity advantage is constantly eroded and needs to be constantly re-created. The creative destruction process, that is, more firm entry and exit and higher portion of innovative young firms, is also fundamental.

In turn, the dynamics of firms and jobs in cities is shaped by the broader characteristics of urban systems. In [Sect. 33.4](#), we highlight major differences between cities in developing countries and more advanced economies. In short, the urban system of many developing countries acts as a brake on economic growth. A major challenge for the countries is the transformation of their urban systems into drivers of economic growth. More specifically, cities in developing countries appear to be far less functionally specialized than cities in more advanced economies. This hampers the dynamism of the largest cities in developing countries which are burdened by many ancillary activities. These activities add to urban crowding without adding to agglomeration benefits. Better infrastructure, in particular better transportation infrastructure, and a reduction in favoritism toward large cities may be a way to remedy these problems. Policies to foster job creations directly may be tempting, but their record in more advanced economies is unsatisfactory. In addition, developing cities also function less efficiently and face challenges that differ from those of cities in more advanced economies. An appropriate management of the transition to full urbanization, a strengthening of urban governance, a reduction in labor market duality, and a reduction or the full elimination of land market duality are key challenges that must be tackled for developing cities to take the full advantage of agglomeration effects and foster aggregate growth.

33.2 Cities, Worker Productivity, and Wages

Cities enjoy a productive advantage over rural areas, and this advantage is larger for larger cities. The positive association between various measures of productivity and urban scale has been repeatedly documented. That larger cities obtain higher scores on many productivity metrics from wages to output per worker, or the total factor

productivity of firms is now beyond doubt. Most of the studies reviewed by Puga (2010) find an elasticity of wages or firm productivity with respect to city employment or urban density between 0.02 and 0.10. As shown by Henderson (2005), these findings also hold widely in cities in developing countries.

More formally, this type of work involves regressing an outcome variable by location on a measure of agglomeration. In the early literature, the typical regression of choice involved using output per worker as dependent variable and city population as explanatory variable. In the early 1990s, authors often employed more indirect strategies and started to use variables such as employment growth or firm creation as outcome measures. More recently, the literature has moved to microdata and returned to more direct outcome measures, namely, the total factor productivity of firms and wages. More precisely, recent studies estimate a regression like

$$\log w_{ic(i)} = \alpha \log Pop_{c(i)} + \eta_{c(i)} + u_i + \varepsilon_{ic(i)}, \quad (33.1)$$

where c denotes cities and i denotes individuals or groups of individuals. The dependent variable is w the wage, and the explanatory variables are $\log Pop$ the log of population as a measure of urban scale, η a city effect (usually proxied through a number of control variables at the city level), and u an individual effect (often proxied through observable individual characteristics). Finally, ε is an error term. The estimated value of the coefficient of interest, α , is usually positive and significant. Similar regressions can be proposed for firm data using measures of firm level productivity and firm characteristics.

After Ciccone and Hall (1996), density has often been favored relative to population since it appears to yield more reliable results. The reason is probably that density-based measures of agglomeration are more robust to zoning idiosyncrasies. For instance treating Washington and Baltimore as one big consolidated metropolitan area or two separate cities makes a big difference to their employment count but only little difference to density.

After asserting this robust statistical association between productivity outcomes and agglomeration, the first question regards whether the estimated coefficient α in the regression described by Eq. (33.1) reflects the causal effect of agglomeration on wages. An examination of Eq. (33.1) reveals three possible sources of bias. They all come from the fact that, as highlighted by the notations in Eq. (33.1) above, the measure of agglomeration is indexed by $c(i)$, that is, the city c is *chosen* by worker i . Ideally, one would like to compare the same workers across the cities that they have chosen and those that they have not chosen. In absence of randomized experiments, this is not possible. Greenstone et al.'s (2010) quasi experiment on “million dollar plants” is what comes closest to this ideal for firms' location choices.

The first source bias is the possible link between city effects (which are not observed directly) and the variable of interest, city population, or density. Put differently, the “quantity of labor” may be endogenous, and it is reasonable to expect workers to go to more productive cities. A possible solution to this problem

is to use instruments for city population or density as Ciccone and Hall (1996). These instruments need to predict current population patterns but must be otherwise uncorrelated with city productivity. Deep historical lags such as population from 200 years ago or soil characteristics can do the job. Studies using this type of approach typically find that correcting for the endogeneity of population has only a mild downward effect on the estimation of the coefficient of interest α .

The second main identification problem in the estimation of Eq. (33.1) regards a possible correlation between the measure of city population and individual effects. That is, the quality of labor may be endogenous, and we expect more productive workers to reside in larger cities. A first possible solution to this problem is to control for an extensive set of individual characteristics. A more drastic solution is to use (whenever possible) the longitudinal dimension of the data and impose worker fixed effects as Combes et al. (2008). The endogenous quality of labor seems to be an important source of bias in the estimation of Eq. (33.1). The estimated value of α is typically reduced by 30–50 % using extensive individual controls or worker effects. This said, one needs to be careful. Imposing worker effects improves the quality of the estimation, but it is not a perfect solution since it assumes that mobility is exogenous.

Related to this last issue, the third source of bias in the estimation of Eq. (33.1) is the possibility of a correlation between the error term and the measure of city population of interest. If, for instance, workers move more easily from large cities to small cities than the opposite in case of a good external wage offer, this will create another source of bias which in this particular situation leads to an underestimate of agglomeration economies. No satisfactory solution to this problem has been proposed so far.

At this point, the conclusion of the agglomeration literature is that there is a causal static effect of cities and urbanization on wages in more advanced economies but that this effect represents only about half the measured association between city population or density and wages (or alternative measures of productivity). The rest of the association between population or density and wages reflects the sorting of more productive workers in larger and denser cities and, to a lesser extent, reverse causality and workers moving to more productive places. Recent investigations that tackle the concerns mentioned above find agglomeration elasticities around 2 %. They thus suggest rather modest static effects of cities on productivity. The literature from developing countries often uses less sophisticated approaches but finds results that are comparable and, if anything, indicative of moderately stronger agglomeration effects.

After questioning its causal aspect, the second key question about the estimation of agglomeration effects regards their sources. When asking about the “sources” of agglomeration, the literature frequently confuses two separate questions. The first is about which markets are affected by these agglomeration effects, and the second is about which mechanisms actually occur. Regarding the “where” question, it is customary to distinguish the markets for (intermediate) goods, the market for labor, and the (absent) market for ideas and knowledge. In terms of mechanisms, we often distinguish between sharing, matching, and learning mechanisms.

“Sharing” is about the many possible benefits from the mutualization of specialized input providers, the diversity of local goods, the division of labor, or the risks. “Matching” is about the greater probability of finding another party such as a worker, an employer, a supplier, or an investor and the greater quality of the match with that party. Finally, “learning” is about the better generation, diffusion, and accumulation of knowledge. The latter set of mechanisms is regularly referred to as knowledge spillovers.

Because of the wide variety of possible mechanisms and the markets where they can take place, the literature that investigates the sources of agglomeration benefits is much more heterogeneous than the literature that attempts to measure the overall benefits from agglomeration. The latter naturally coalesces around the estimation of Eq. (33.1).

First, there is a diversity of work which provides evidence of an association between some aspect of agglomeration such as a particular mechanism or market and measures of agglomeration such as city size. Let us take only a few recent examples (see Puga 2010, for a more exhaustive discussion). Taken together, these studies are suggestive that many of the agglomeration mechanisms described by the theoretical literature are at work in a variety of markets.

This conclusion must be taken cautiously, however. Establishing the direction of causality in this type of work is even harder here than when attempting to measure the overall effects of agglomeration. To understand this point and the pitfalls associated with this type of work, let us use the analysis of Charlot and Duranton (2004) on workplace communication. They show that communication is associated positively with city size and with wages. This leads them to conclude that communication spillovers could account for up to a quarter of agglomeration benefits. However, this finding could be explained in part by the greater sorting of good communicators in larger cities. This is the equivalent of the quality-of-labor bias discussed above. This worry can be reduced by comparing movers and stayers in cities as Charlot and Duranton (2004) do. It is difficult to eliminate it entirely though. In addition, one also needs to show that greater communication in cities is not the by-product of another agglomeration force. Workers in larger cities may communicate more because firms outsource more of their output. This requires some coordination. In such a case, the real source of agglomeration benefits may be input–output linkages, not communication spillovers. To go round this problem, Charlot and Duranton (2004), who use rich firm level data, suggest instrumenting workplace communication by measures of organizational changes such as a flattening of the hierarchy. These changes typically increase the need for horizontal communication. This type of instrument is nonetheless valid only if changes in organization are unrelated to other sources of agglomeration benefits such as labor pooling or input–output linkages. That firm reorganization affects worker communication behavior but has no direct effect on recruiting practices, or outsourcing is plausible but not certain. More generally, studies that focus on one particular source of agglomeration face a major missing variable problem: The other sources of agglomeration are absent from the regression even though they are expected to be correlated with both wages (or other productivity measures) and measures of agglomeration such as city size.

Given how difficult it is to measure many aspects of agglomeration and given also that the list of possible agglomeration sources is open, considering all sources of agglomeration in one regression is not a feasible option. A more reasonable path forward is, following Ellison et al. (2010), to consider several classes of agglomeration sources in the same approach. Ellison et al. (2010) assess how much labor pooling, input–output linkages, and spillovers account for co-agglomeration between industries in the USA. They use a measure of industry co-agglomeration and find more co-agglomeration among (i) industries that buy from each other, (ii) industries that use a similar workforce, and (iii) industries that share a common scientific base. To reduce the possibility that co-agglomerated industries end up buying from each other or using similar workers because of their proximity, they instrument their US measures of input–output linkages and labor pooling using corresponding UK data. Of course, if the biases are the same in the UK as in the USA, these instruments are of limited value. Another caveat is that input–output linkages are possibly more easily measured using input–output matrices than spillovers using patent citations. This can also lead to biased estimates since a positive correlation with both linkages and spillovers is likely to be picked up mainly by the better-measured linkage variable. This said, Ellison et al. (2010) confirm that the three motives for agglomeration they consider are at play with input–output linkages playing a more important role.

Even if we abstract from the uncertainty around those results, the notion that several mechanisms, each operating in several markets, contribute to agglomeration benefits is problematic for policy. At their heart, agglomeration benefits rely on market failures associated with the existence of small indivisibilities with sharing mechanisms, thick market effects with matching mechanisms, and uncompensated knowledge transfers with learning mechanisms. That is, there are possibly many market failures at play in many markets. In turn, this implies that there may be no hope of fostering agglomeration economies through a small number of simple policy prescriptions.

Before broadening the discussion, there are four further features of agglomeration that have implications for workers and jobs in cities.

The first is the issue of the sectoral scope of agglomeration and whether agglomeration effects accrue mostly within or across sectors. Agglomeration effects within sectors are referred to as localization economies and between sectors as urbanization economies. When estimating a more general version of Eq. (33.1) that accounts for both city size or density and the degree of same sector specialization, extant research has found evidence of both localization and urbanization effects. There are two interesting nuances. The first is the presence of significant heterogeneity across industries. This heterogeneity follows an interesting pattern as it appears that more technologically advanced industries benefit more from urbanization economies whereas more mature industries benefit more from localization economies. Second, the calculations of Combes et al. (2008) indicate that in France the benefits from localization economies are smaller than those of urbanization economies and mostly uncorrelated with local wages. Put differently, increased local specialization has only small benefits and does not contribute to making workers richer.

The second extra feature of agglomeration is the notion that not all workers benefit equally from urban scale. Equation (33.1) estimates an “average” agglomeration effect. As highlighted by, among others, Glaeser and Resseger (2010), agglomeration effects appear stronger for more educated workers in the USA. Higher returns in larger cities should in turn provide stronger incentives to more skilled workers to locate there. Hence, these results are consistent with the well-documented fact that workers in larger cities in more advanced economies tend to be more educated and better skilled (e.g., Combes et al. 2008).

Next, while not all workers benefit equally to agglomeration effects, it also appears that not all workers contribute equally to these effects either. There is a large literature on human capital externalities suggesting that workers enjoy higher wages when surrounded by more educated workers. Estimates of external returns to education are typically between 50 % and 100 % the corresponding estimates of private returns to education, in particular for university graduates. These findings are robust to a number of estimation concerns and suggestive of large effects. It is beyond the scope of this chapter to review this literature extensively. See instead Moretti (2004) for an in-depth survey.

Finally, there is also emerging evidence from US and European data that wage growth also depends on city size/density. To show this, one can estimate a regression along the lines of Eq. (33.1) but use wages in first difference instead of in levels as dependent variable:

$$\Delta_{t+1,t} \log w_{ic(i)} = \alpha \log Pop_{c(i)t} + \eta_{c(i)} + u_i + \varepsilon_{ic(i)t} \quad (33.2)$$

where Δ is used to note time differences between t and $t + 1$. Among a number of papers, De la Roca and Puga (2012) confirm that wage growth is stronger in larger cities.

Because the structure of Eq. (33.2) is the same as that of Eq. (33.1) for the static estimation of agglomeration economies, it suffers from the same drawbacks. First, the association between wage growth and agglomeration could be explained by the sorting of workers with faster wage growth in larger cities. This could occur because “fast learner” tends to locate in larger cities or because the wage of workers who are predominantly located in larger cities (such as more educated workers) tends to increase faster. Following the same sort of fixed effect strategy described above and applying that to a regression like Eq. (33.2), Freedman (2008) nonetheless shows that this type of result holds even after controlling for the fact that some workers may experience higher wage growth independently of their location.

Although the result that wages grow faster in cities is frequently interpreted as evidence about faster learning in cities and knowledge spillovers, the mechanisms that drive it are unclear. Just like regressing wages in levels on a measure of urban scale in Eq. (33.1) does not tell us anything about the sources of static agglomeration economies, regressing wage growth on urban scale in Eq. (33.2) is equally uninformative about the sources of agglomeration dynamics. Interestingly, Wheeler (2008) shows that young workers tend to change job more often in larger cities, while the opposite holds for old workers. This type of evolution is consistent with

a matching model where workers can find their “ideal match” faster in larger cities and then stick to it. Such mechanism could explain both faster wage growth and eventually higher wages in larger cities.

Evidence about learning in cities can come from the fact that workers retain some benefits from agglomeration after they leave their city. De la Roca and Puga (2012) confirm this on Spanish data. Their findings suggest the existence of both a level effect of cities on wages (of the same magnitude as those discussed above) and a dynamic effect. Over the long run, workers in large cities seem to gain about as much from both effects.

To sum up, this discussion of agglomeration economies which focuses mainly on workers and jobs reaches a number of interesting conclusions. First, larger cities make workers more productive. There is both a static and a dynamic component to these gains. A static elasticity of wages with respect to city population of 0.03 implies that a worker receives a 23 % higher wage when moving from a tiny city with population 5,000 to a large metropolis with a population of five million. Over time, dynamic effects could make this urban premium twice as large. While long-run gains close to 50 % are not miraculous, they are nonetheless sizeable.

In terms of policy implications, the temptation to “foster agglomeration effects” should be resisted. We are too far from knowing enough about the sources of agglomeration to implement any meaningful policy, not to mention the great heterogeneity in who gains from and who contributes to agglomeration gains. It remains nonetheless that the economic gains from urbanization are significant and urbanization should be embraced rather than resisted.

33.3 Firm Dynamics Within Cities

This higher productivity of jobs in cities is only one facet of the issue. Jobs are usually viewed as a veil when we model production in theoretical models. In practice, higher labor productivity is associated with doing different things and doing them differently. That is, to receive higher wages, workers need “better jobs.” Firm dynamics is often the vector of these changes. More specifically, let us examine several aspects of firm dynamics in cities: innovation, firm creation and growth, and factor allocation and reallocation across firms.

Starting with innovation, the first salient feature of the geography of innovative activity is that research and innovation is much more concentrated than production in most industries. Interestingly, this tendency seems particularly strong for industries that are more intensive in skilled labor and in research and development. It is also the case that this concentration of research and development typically takes place in large metropolitan areas.

These location patterns for innovative activity are consistent with the notion that cities have a positive effect on innovation just like they have on wages. More direct evidence can be found in Feldman and Audretsch (1999) and Carlino et al. (2007). To measure innovation, Feldman and Audretsch (1999) make a count of all new product innovations in US metropolitan areas for a broad set of technologies and

sectors in 1982. They find no evidence of urban scale effects but find that same sector specialization is strongly negatively associated with innovation whereas a diversity of employment in technologically related industries is strongly positively associated with innovation. They also find strong positive innovation effects associated with the presence of smaller establishments.

Using the number of patents per capita as dependent variable, Carlino et al. (2007) find evidence of strong agglomeration effects for innovative activity. Their estimate of the elasticity of patenting per capita with respect to employment density is 0.2. This is several times the estimates reported above for the corresponding elasticity of wages. Interestingly, Carlino et al. (2007) also find that this elasticity of innovation with respect to employment density or population size is not constant across the urban hierarchy. Patenting per capita appears to peak at around 5,700 jobs per square kilometer or a city population size slightly below a million.

While this evidence is highly suggestive that cities affect innovation, there is, to the best of our knowledge, no work which focuses on the effects of innovative activity in cities such as its effects on urban growth. Regressing urban population growth on innovative activity would raise some obvious identification concerns. In addition, simple theoretical argument suggests that the effect of innovation on urban growth need not be positive. Obviously, product innovation in the form of either an entirely new product or the capture of an established product from another location is expected to add to a city's employment. However, process innovation within a city can cut both ways. Employment will increase with process innovation only if greater productive efficiency and lower prices lead to a more than proportional increase in demand. In the opposite case, process innovation will imply a contraction of local employment. Remarkably, Carlino et al. (2007) show that Rochester, Buffalo, Cleveland, St Louis, and Detroit are all highly innovative cities. This suggests that, to some extent, the demise of these cities may be attributed to the fact that labor productivity increased much faster than demand in their industries.

Finally, innovative activity appears to change the nature of jobs in the cities where it takes place. As shown by Lin (2011), cities that patent more tend to have a greater proportion of what he labels "new work," that is, jobs that did not exist a few years before. New work is also fostered by a greater proportion of educated workers and a diversity of industries, two other attributes of large cities.

To conclude on the links between innovation and cities, extant literature supports the notion that cities affect innovation either because of their sheer population size or because of the (diverse) structure of their production activities. The evidence about the effect of innovation on cities is more complex. Innovation *within a given city* affects the proportion of workers in new work. Other effects are either ambiguous or poorly documented. As we show below, further insights about the effects of innovation on cities can be gained by looking across cities.

Entrepreneurship is also closely associated with cities in several ways. First, cities affect entrepreneurship just like they affect wages and innovation. In a comprehensive analysis of the determinants of employment in new manufacturing start-ups across sectors in US cities, Glaeser and Kerr (2009) generate a rich harvest of facts. The first is the existence of scale economies. As a city grows larger,

employment in new start-ups in this city increases more than proportionately. Depending on their specification, Glaeser and Kerr (2009) find an elasticity of employment in new start-ups per capita with respect to city scale between 0.07 and 0.22. City population, city-industry employment, and sector effects explain around 80 % of the variation in start-up employment across cities and sectors.

Glaeser and Kerr (2009) also find that the presence of many small suppliers has a strong effect on employment in start-ups. In addition, they also find evidence of mild Marshallian effects associated with input–output linkages, labor market pooling, and spillovers. Finally, city demographics only has a limited explanatory power just like their measure of “entrepreneurial culture.”

The other key feature about the supply of entrepreneurs is that there is a strong local bias in entrepreneurship. Entrepreneurs tend to create their start-up in the place where they were born and/or where they have lived and worked before becoming entrepreneurs. This important fact has been documented by Figueiredo et al. (2002) for Portugal and Michelacci and Silva (2007) for Italy and the USA. This finding has been confirmed by several other studies in developed economies. Figueiredo et al. (2002) also show that when entrepreneurs chose a new location, this choice is strongly governed by agglomeration economies and a proximity to large cities.

After looking at the urban determinants of entrepreneurship, we now turn to the effects of entrepreneurship on their cities. It has been shown repeatedly that entrepreneurship plays a key role in urban evolutions. The key fact here is that growth in a city and sector over a period of time is strongly correlated with the presence of small establishments in that city and sector at the beginning of the period. This fact was first documented by Glaeser et al. (1992) and has been confirmed for other countries and time periods by many other studies.

Just like with many of the correlations discussed above, the strong link between small firms and employment growth raises a key identification concern about the direction of causality. However, this issue has been neglected by the literature. This is perhaps because the standard regression uses growth over a period as dependent variable and establishment size *at the beginning of the period* as explanatory variable. However, using a predetermined variable as explanatory variable in a regression does not guarantee its exogeneity. Local entrepreneurs could enter in large numbers in a city and sector if they foresee strong future demand. That expectations of future growth should trigger entry today is only natural. That is the nature of business.

To resolve this identification problem, it is difficult to think of instruments that would predict establishment size in a city and sector but be otherwise uncorrelated with subsequent growth. To clarify the meaning of the relationship between small establishments and high subsequent growth, Glaeser et al. (2010) do something quite different. They look at whether the presence of many small firms in a city and sector is driven by the demand for entrepreneurship or its supply. To the extent that the demand for entrepreneurship can be captured by higher sales per worker, this does not appear to be the case. They also find limited evidence about the importance of lower labor costs or entrepreneurs sorting into high amenity cities. They find

stronger evidence about the importance of the proportion of university graduates (particularly in more skilled industries), but that still does not explain away the effect of having lots of small establishments. While still preliminary, this type of evidence points at some unspecified supply effects. More entrepreneurial cities happen to have a greater supply of entrepreneurs, and the literature has thus far been unable to trace this further.

Turning finally to factor allocation and reallocation, the literature that examines these issues makes two important claims. The first is that a large fraction of productivity growth at the country level can be accounted for by the reallocation of factors from less productive to more productive firms. A large share of productivity growth can be accounted for by a churning process where low-productivity firms are replaced by new and more productive start-ups. These important findings have been confirmed for many countries (Bartelsman et al. 2004).

The second important claim made by the reallocation literature is that “misallocation” can account for a large share of existing productivity differences across countries. To understand this point better, consider the influential work of Hsieh and Klenow (2009). They first note that, in equilibrium, the marginal product should be equalized across firms. If the demand for the varieties produced by firms has a constant elasticity of substitution, this implies an equalization of the product of their price by their “true productivity” (which is the ability of firms to produce output from inputs). This – price times true productivity – product is what is estimated as “total factor productivity” in most productivity exercises. We may call this second quantity “apparent productivity.” Obviously, the firms’ apparent productivities are never equalized in real data. Hsieh and Klenow (2009) interpret this as evidence of factor misallocation. Taking the highly dispersed distribution of manufacturing productivity in China and India, they calculate very large potential costs from such misallocation. Acknowledging that a perfectly efficient allocation may be impossible, they compute that the productivity gains for manufacturing in China and India would still be of about 50 % if their level of misallocation could be reduced to that observed in the USA.

To the best of our knowledge, there is no study that would attempt to relate greater churning/reallocation at the firm level and higher productivity growth at the urban level. However, there is a strong suspicion that larger cities should exhibit more churning. This is because, as already argued, larger cities are more innovative, experience more entry and exit, and have a greater fraction of their workforce in “new work.” At the same time, there is no indication that this greater amount of churning in larger cities is associated with higher productivity growth in those cities unlike what occurs at the country level.

We actually know little about productivity growth in cities. According to Lin (2011), the greater proportion of workers employed in new work in larger cities is not associated with faster productivity growth. In a rare study of the broader determinants of productivity growth in Italian cities, Cingano and Schivardi (2004) highlight the importance of both specialization and employment size. But given that specialization and employment size are negatively correlated, their positive effects arguably cancel out. Hence, more churning does not appear to lead to faster productivity growth in cities.

To confirm this conclusion, note that workers are somewhat mobile across cities. Then more churning associated with faster productivity growth in larger cities should imply a divergence in population growth rates. There is no evidence of such divergence. This lack of result regarding the link between churning and productivity should not be taken as negative evidence against the reallocation literature. As argued in the next section, it is possible that reallocation does not take place within cities but also across cities.

Turning to the second claim about misallocation, Combes et al. (2011) show that the distribution of firm productivity is unambiguously more dispersed in larger cities in France. In the framework of Hsieh and Klenow (2009), that would be interpreted as greater misallocation in larger cities. This seems hard to believe. The evidence about static agglomeration effects discussed above is instead best interpreted as agglomeration economies leading to a better allocation of resources (in a broad sense) in larger cities. When performing a productivity decomposition, Combes et al. (2011) find a similar covariance between establishment size and productivity in large and small cities which suggest a similar level of efficiency in the allocation of factors to firms across cities of all sizes.

To sum up, the evidence about firm dynamics and cities presented in this section is puzzling. Larger cities seem to be more innovative, be more entrepreneurial, experience more churning and reallocation, and generally enjoy a greater “economic dynamism.” At the same time, they do not appear to enjoy most of the benefits associated with such dynamism since neither productivity nor population appears to increase faster in larger cities. Of course, these conclusions need to be taken cautiously given the paucity of study, including their complete absence for cities in developing countries.

33.4 City Functionality, Urban Systems, and Policies

The answer to the apparent puzzle raised above is that when thinking about economic growth, it is wrong to think of cities as self-contained units. Cities are best viewed as small open economies which interact a lot with other cities and rural areas. They are part of an “urban system.” This implies that innovation, churning, and reallocation are best studied across the entire system of cities.

Starting with innovation, recall that larger cities offer many advantages for both product and process innovation. More specifically, as highlighted by many, cities favor the circulation and cross-fertilization of ideas. This naturally leads to more product innovations, and this is consistent with the evidence of Feldman and Audretsch (1999) discussed above. For process innovation, Duranton and Puga (2001) underscore the greater availability of intermediate goods in large cities which allows firms to proceed through trial and error at a faster pace. Put differently, the greater ability of larger cities to innovate may just be another manifestation of agglomeration economies. The key difference with many static aspects of agglomeration economies such as thicker local labor markets is that, with dynamic effects, co-location is not needed all the time. More precisely, spillovers may matter to develop an innovation, but after this is done, co-location is no longer needed.

Quite the opposite, larger cities are more expensive places to produce. After the dynamic benefits from agglomeration have been exploited, it can make sense for firms to relocate. Often, the entire firm does not need to relocate since it is only the production of particular products that is concerned.

Patterns of establishment relocations in France are highly consistent with this type of product cycle. As shown by Duranton and Puga (2001), about 75 % of French establishments that relocate do so from a city with above-median diversity to a city with below-median diversity and above-median specialization in the same sector. In addition, as documented by Fujita and Ishii (1998), large Japanese multinationals in the electronic sector produce their newest products in “trial” plants near Tokyo and Osaka. Less recent products are produced in rural locations in Japan while even older generations of their products are manufactured in less advanced countries in Asia. Hence, as their products mature, firms still search for agglomeration economies but will put a greater weight on the benefits of specialization. Large cities act as nurseries for new goods and new products. Once mature, new goods and products are best produced in more specialized places.

Cities are also specialized by sector. However, this tendency, while still present in the data, has diminished over time as documented by Duranton and Puga (2005). The same authors also document a rise in the functional specialization of cities with the emergence of cities specialized into management-type functions, whereas others specialize more into production activities. This rise in functional specialization is rationalized by Duranton and Puga (2005) in a model where lower communication costs make it easier for firms to separate management from production. Since these activities benefit from very different types of agglomeration economies, such separation is beneficial, provided the cost of separating activities is low enough. In turn, this separation of activities reinforces the functional specialization of cities.

These multiple dimensions of specialization are part of well-functioning urban systems in more advanced countries. Adding to this, the notion of cities being specialized by functions and activities is not static. The process of continuous location and relocation of economic activity is a crucial aspect of the growth of those activities. To take a simple example, when George Eastman developed a new revolutionary technology in the photographic industry in Rochester, the latter relocated from New York to Rochester. Then, much later, as the technology developed by Eastman got itself superseded by the digital revolution, Rochester lost its status of capital of the photographic industry.

That different cities specialize into different functions and are able to change their specialization after negative shocks presupposes a fair amount of “mobility” across cities. The first important dimension of mobility regards goods and services. It would make little sense for cities to narrowly specialize in an activity if its output cannot be exported. Continuously changing patterns of specialization also require labor mobility. For instance, Kerr (2010) documents that after “breakthrough” innovations, more innovations tend to take place in the same location for the same technology. This growth in patenting, in turn, depends on the mobility of scientists and engineers. Interestingly, the adjustment appears faster for technologies that depend more heavily on immigrant inventors who are more mobile.

While the foregoing discussion describes well what happens within the urban system of more advanced economies, it is a far less resemblant depiction of the situation of cities in developing countries. For instance, most very large cities in developing countries are still major manufacturing centers, whereas manufacturing production is mostly absent from the largest cities of Europe and North America. This lack of urban differentiation may be at the root of the problem. Urban systems in developing countries may be much less efficient than in more advanced countries because cities are much less differentiated in terms of functions.

More specifically, this lack of differentiation in urban functionality may hamper the dynamism of cities in developing countries. The largest cities there are burdened by many ancillary activities such as basic manufacturing and call centers. These add to urban crowding without adding to agglomeration benefits. On the other hand, smaller cities in developing countries often lag far behind, and getting some of these ancillary activities would be crucial for their development.

This said, a lack of well-functioning urban systems – however important (and neglected in urban policy) – is not the only cause for the lower efficiency of cities in developing countries relative to their counterparts in advanced economies. Nonurban factors such as weak national institutions and poor technology certainly play a role. Urban factors which hinder the functional differentiation of cities also have a direct negative effect on the efficiency of cities. For instance, as we discuss below, high transportation costs limit the specialization of cities by reducing their ability to trade. At the same time, even if we abstract from these effects, high transportation costs also affect the price of goods purchased by local consumers and reduce market access for local producers.

In the rest of this section, we examine a number of urban factors that both reduce the efficiency of the urban system as well as the efficiency of cities directly. Cities in developing countries are often acting as a brake on growth, whereas they should be a key driver of economic development.

The first key difference between cities in developing and more advanced countries regards the functioning of their labor market. In most developing countries, there is a well-known duality in the labor market which usually comprises a large informal sector alongside the formal sector. Aside from its detrimental implications for workers in the informal sector, this duality hinders urban development in several ways. First, it has been accused of inducing too much migration toward the largest cities where most of the formal sector is located. Duality may also limit mobility across cities since jobs in the informal sector tend to be filled by word-of-mouth through social connections which are missing to newcomers. High barriers to “good” jobs in the formal sector may also hold back the incentives of workers to improve their skills locally and thus limit the scope of agglomeration benefits.

To mitigate the effects of labor market duality, three broad types of policies can be envisioned. The first is to improve the working of labor markets. While this objective is certainly laudable, a discussion of this class of policies would certainly go beyond the scope of this chapter.

The second type of policy is to foster local job creation through “place-based” policies. Such policies typically involve tax exemptions or subsidies associated

with job creation within well-defined (and often tightly circumscribed) areas. These tools are frequently used to try to reduce the unemployment rate of the residents of poor areas in more advanced economies. While the labor market failures in developed and developing countries differ and the scale at which such policies might be implemented in developing countries may be much broader than poor neighborhoods of “rich” cities, there may be useful lessons to learn from the recent North American and European literature evaluating those policies. Simply put, the general record of place-based policies is in doubt. Detailed evaluations of particular policies are usually negative (Glaeser and Gottlieb 2008).

The third class of policies attempts to foster job creations in a particular locality by helping firms in a given sector. These policies are usually referred to as “cluster” policies and follow from the work of Michael Porter (1990). They often entail the development of subsidized supportive institutions and infrastructure using public subsidies and various types of fiscal incentives. The review of the literature in Puga (2010) implies negative conclusions about the possible benefits of cluster policies.

The second key difference between cities in developing and more advanced countries regards the functioning of their land market. Like labor markets, land markets in developing cities are characterized by a duality between land used with appropriate property titles and leases and squatted land. Recent empirical research has focused on the effects of the lack of effective, formal property titles which could prevent residents of squatter settlements from using their house as collateral. Informal land markets may thus be a major barrier to enterprise development. The empirical evidence about the relaxation of credit constraints associated with “titling” policies is weak. Recent work points instead to increases in labor supply (Field 2007) and to the adoption of more middle-class values and attitudes (Di Tella et al. 2007). While this evidence about titling policies is relatively optimistic about the merits of such policies, it must be noted that the existing literature focuses nearly exclusively on residential land. The extent of land illegality for commercial land (from illegal street vendors to squatter manufacturing) is poorly measured, and the solutions are not well developed.

The third key difference between cities in developing and more advanced countries regards infrastructure, particularly the road infrastructure. Two strands of research need to be distinguished here. The first finds its roots in international trade and focuses on the estimation of the effect of “market potential” variables. The market potential of a city is usually computed as the sum of the income (or population) of other cities weighted by their inverse distance to the city under consideration. Assuming transportation costs and other trade frictions associated with distance, many models of international and interregional trade generate the prediction that a location’s income will be determined by its market access (Krugman 1991). The literature offers strong empirical support regarding the importance of market access for cities in developing countries (Henderson 2005).

The second strand of literature focuses more closely on the effects of infrastructure. Baum-Snow’s (2007) pioneering work finds that the construction of the interstate highway system was a major impetus behind the suburbanization of US cities. Duranton and Turner (2012) also find that more kilometers of interstate

highways in US metropolitan areas in the early 1980s led to faster population growth over the subsequent 20 years.

This type of approach is also being applied to developing countries. In a remarkable piece of work, Donaldson (2010) documents the effects of the construction of India's railroad network by its colonial power. He shows that railroads increased trade and reduced price differences across regions. Even more importantly railroads increased real incomes and welfare. To minimize identification problems, he compares the network that was built to other networks that were considered but never developed.

In line with some of the arguments advanced above about the importance of transportation infrastructure for the decentralization of manufacturing activity away from large metropolises, Baum-Snow et al. (2011) underscore the importance of railroads in the decentralization of manufacturing production in China.

Storeygard (2011) provides evidence about the importance of inter-city transportation costs for inland African cities. Using new roads data for Africa and satellite data ("lights at night") to estimate economic activity, he assesses the effect of higher transportation costs. To circumvent the endogeneity of transportation costs (roads may be built to access growing cities), he uses arguably exogenous variations in oil prices. He finds an elasticity of economic activity with respect to transportation costs of about -0.2 .

All these findings are suggestive of the profound and long-lasting effects of major transportation infrastructure. One needs to keep in mind nonetheless that major transportation networks are extremely costly investments.

The last key difference between cities in developing and more advanced countries regards the effects of the favoritism by governments of the largest cities. While the reasons for primate city favoritism are still debated (Henderson 2005), there is little doubt that such favoritism takes place in many different ways. As argued in Henderson (2005), primate city favoritism harms the favored primate city by making it bigger than it should be. It also harms smaller cities which are, in effect, heavily taxed. The gap that is created between the primate city and other cities may also have negative dynamic effects since for most educated workers there is nowhere to go except stay in this primate city. As a result this may reduce the circulation of knowledge across cities. Reducing primate city favoritism and providing smaller cities with better local public goods (including education and health) are certainly a big part of any solution.

33.5 Conclusions

For individual workers, cities in developing countries appear to bring significant benefits both in the short run and in the long run. However, when taking a broader look, the urban system of developing countries appears to involve far less functional differentiation across cities than in more advanced economies. Such differentiation with different cities playing different roles in the urban system is important for the process of growth and development to proceed smoothly.

Larger cities innovate and manage. Smaller cities often produce a narrow range of goods. Having larger cities do everything like they often in developing countries reduces their dynamism and holds back small cities which remain stagnant.

A variety of policies can be envisioned to solve this problem. The three more promising areas are general policies to improve the functioning of labor markets, ending primate city favoritism, and development of major infrastructure to connect cities.

References

- Bartelsman E, Haltiwanger J, Scarpetta S (2004) Microeconomic evidence of creative destruction in industrial and developing countries. University of Maryland, Mimeo
- Baum-Snow N (2007) Did highways cause suburbanization? *Q J Econ* 122(2):775–805
- Baum-Snow N, Brandt L, Henderson JV, Turner MA, Zhang Q (2011) Roads, railways and decentralization of Chinese cities. Brown University (Processed)
- Carlino GA, Chatterjee S, Hunt RM (2007) Urban density and the rate of invention. *J Urban Econ* 61(3):389–419
- Charlot S, Duranton G (2004) Communication externalities in cities. *J Urban Econ* 56(3):581–613
- Ciccone A, Hall RE (1996) Productivity and the density of economic activity. *Am Econ Rev* 86(1):54–70
- Cingano F, Schiavardi F (2004) Identifying the sources of local productivity growth. *J Eur Econ Assoc* 2(4):720–742
- Combes P-P, Duranton G, Gobillon L (2008) Spatial wage disparities: Sorting matters! *J Urban Econ* 63(2):723–742
- Combes P-P, Duranton G, Gobillon L, Puga D, Roux S (2011) The productivity advantages of large cities: distinguishing agglomeration from firm selection. *Econometrica* 80(6):2543–2594
- De la Roca J, Puga D (2012) The dynamic earnings premium of dense cities. CEMFI and IMDEA Sociale Sciences (Processed)
- Di Tella R, Galliani S, Schargrodsky E (2007) The formation of beliefs: evidence from the allocation of land titles to squatters. *Q J Econ* 122(1):209–241
- Donaldson D (2010) Railroads of the Raj: estimating the impact of transportation infrastructure. MIT (Processed)
- Duranton G, Puga D (2001) Nursery cities: urban diversity, process innovation, and the life cycle of products. *Am Econ Rev* 91(5):1454–1477
- Duranton G, Puga D (2005) From sectoral to functional urban specialisation. *J Urban Econ* 57(2):343–370
- Duranton G, Turner MA (2012) Urban growth and transportation. *R Econ Stud* 79(4):1407–1440
- Ellison G, Glaeser EL, Kerr WR (2010) What causes industry agglomeration? Evidence from coagglomeration patterns. *Am Econ Rev* 100(3):1195–1213
- Feldman MP, Audretsch DB (1999) Innovation in cities: Science-based diversity, specialization and localized competition. *Eur Econ Rev* 43(2):409–429
- Field E (2007) Entitled to work: urban property rights and labor supply in Peru. *Q J Econ* 122(4):1561–1602
- Figueiredo O, Guimarães P, Woodward D (2002) Home-field advantage: location decisions of Portuguese entrepreneurs. *J Urban Econ* 52(2):341–361
- Freedman M (2008) Job hopping, earnings dynamics, and industrial agglomeration in the software publishing industry. *J Urban Econ* 64(3):590–600
- Fujita M, Ishii R (1998) Global location behavior and organizational dynamics of Japanese electronics firms and their impact on regional economies. In: Chandler AD Jr, Hagström P, Sölvell Ö (eds) *The dynamic firm: the role of technology, strategy, organization and regions*. Oxford University Press, Oxford, pp 343–383

- Glaeser EL, Gottlieb JD (2008) The economics of place-making policies. *Brook Pap Econ Act* 1:155–253
- Glaeser EL, Kerr WR (2009) Local industrial conditions and entrepreneurship: how much of the spatial distribution can we explain? *J Econ Manag Strategy* 18(3):623–663
- Glaeser EL, Resseger MR (2010) The complementarity between cities and skills. *J Reg Sci* 50(1):221–244
- Glaeser EL, Kallal H, Scheinkman JA, Shleifer A (1992) Growth in cities. *J Polit Econ* 100(6):1126–1152
- Glaeser EL, Kerr WR, Ponzetto GAM (2010) Clusters of entrepreneurship. *J Urban Econ* 67(1):150–168
- Greenstone M, Hornbeck R, Moretti E (2010) Identifying agglomeration spillovers: evidence from winners and losers of large plants openings. *J Polit Econ* 118(3):536–598
- Henderson JV (2005) Urbanization and growth. In: Aghion P, Durlauf SN (eds) *Handbook of economic growth*, vol 1B. North-Holland, Amsterdam, pp 1543–1591
- Hsieh C-T, Klenow PJ (2009) Misallocation and manufacturing TFP in China and India. *Q J Econ* 124(4):1403–1448
- Kerr WR (2010) Breakthrough inventions and migrating clusters of innovation. *J Urban Econ* 67(1):46–60
- Krugman PR (1991) Increasing returns and economic geography. *J Polit Econ* 99(3):484–499
- Lin J (2011) Technological adaptation, cities, and new work. *Rev Econ Stat* 93(2):554–574
- Michelacci C, Silva O (2007) Why so many local entrepreneurs? *Rev Econ Stat* 89(4):615–633
- Moretti E (2004) Human capital externalities in cities. In: Henderson V, Thisse J-F (eds) *Handbook of regional and urban economics*, vol 4. North-Holland, Amsterdam, pp 2243–2291
- Porter ME (1990) *The competitive advantage of nations*. Free Press, New York
- Puga D (2010) The magnitude and causes of agglomeration economies. *J Reg Sci* 50(1):203–219
- Storeygard A (2011) Farther on down the road: transport costs, trade and urban growth in sub-Saharan Africa. Brown University (Processed)
- Wheeler CH (2008) Local market scale and the pattern of job changes among young men. *Reg Sci Urban Econ* 38(2):101–118

Changes in Economic Geography Theory and the Dynamics of Technological Change

34

Riccardo Crescenzi

Contents

34.1	Introduction	650
34.2	The Linear Model of Innovation: The A-Spatial Benchmark	651
34.3	Physical “Distance” Between Innovative Agents and Knowledge Flows	653
34.4	Innovative Agents “in Context”: Local Specialization Patterns and Institutions	655
34.4.1	Economic Places: Industrial Specialization	655
34.4.2	Relational-Institutional Places	656
34.5	Bringing Different Approaches Together: Nonspatial Proximities and “Integrated” Frameworks	658
34.6	Conclusions	662
	References	664

Abstract

This chapter looks at the recent developments in economic geography theory and sets out to shed light on its contribution to the understanding of the dynamics of technological change. The replacement of the linear model with more sophisticated conceptualizations of the process of innovation has made it possible to account for persistent disparities in innovative performance across space and has motivated researchers to incorporate the role of space and places in the analysis of innovation processes. From the physical-metrical approach of geography as distance to the emphasis on specialization and diversification patterns (geography as economic place), institutional-relational factors, nonspatial proximities, and “integrated” frameworks, economic geography theory has substantially

R. Crescenzi

London School of Economics, London, UK

e-mail: r.crescenzi@lse.ac.uk

evolved in terms of its contribution to the understanding of technological dynamics with significant implications for the rationale, design, and implementation of innovation policies.

34.1 Introduction

In an increasingly globalized world of intensified competition with ever-shorter product life cycles, new technologies and innovation are key determinants of regional and national competitiveness. This is certainly good news for developing and emerging countries and regions: economic performance can be boosted by stronger indigenous innovative capabilities but also by better accessibility to external knowledge. New windows of opportunity are being opened by innovation and technological change for new actors to emerge in the international technological competition arena. However, a large body of empirical evidence suggests that these opportunities are far from “universal”: knowledge generation and absorption are highly localized, and diffusion follows very complex (and ever changing) patterns. In both developing and developed countries, a small number of “hot spots” are pushing the technological frontier forward, followed by a set of emerging second-tier “imitative systems” and a large number of territories that exhibit little innovative dynamism and only marginal benefits from technological opportunities. Innovation is certainly spreading both internationally – as suggested, for example, by the success of China and India – and “nationally” with new territories gaining momentum in the “new” member states of the EU but only in a very circumscribed set of new suitable “locations.” This is true in Europe and the United States where around 70 % of total patenting remains concentrated in the twenty most innovative regions (Crescenzi et al. 2007) but also in China and India where these concentration patterns are even more significant.

Rather than waning, such spatial innovation disparities are increasing in both developed and developing countries, shattering hopes that rapid progress in information and communication technologies (ICT) and the dismantling of barriers to the movement of labor and capital can automatically decouple innovative performance from previous localized patterns of technological accumulation and contextual socio-institutional and geographical conditions. Conversely, the spatial concentration of knowledge generation in a few leading “hot spots” boosts their attractiveness for inward investment in innovative activities, further reinforcing the localization of the key nodes of “global” knowledge networks generated by the mobility of both capital (e.g., by multinational firms and their internal connections) and skilled labor (e.g., diasporic communities), generating a cumulative self-reinforcing process.

Technological change and innovation – with their capability to generate new economic opportunities – are features of cities, clusters, and regions whose contribution toward national and global systems and networks is highly asymmetric. This, therefore, calls for appropriate frameworks of understanding able to capture the two-way nexus between geography and innovation. Coherently with this

perspective, this chapter aims to critically review the existing literature on territorial innovation dynamics in order to shed light on how progressively more sophisticated conceptualizations of the role of geography in innovation dynamics have been developed, and how they can address the complexity of the “real” world processes discussed above in a more effective manner.

When looking at how the literature has conceptualized the economic geography of innovation dynamics, it is possible to identify four major streams of literature:

- a. Being based on physical-metric space, the first stream of literature has analyzed the role of *physical distance* between innovative agents in shaping their innovative capabilities.
- b. The second stream, instead, has focused on geography as an “*economic place*,” looking at how local sectoral and functional specialization patterns shape the generation of innovation.
- c. The third set of contributions has concentrated on *institutional-relational places*, looking at the impact on innovation of the rules and patterns shaping the interactions between innovative agents in a given locality.
- d. The final set of academic works has developed the idea that economic and institutional-relational processes can be de-coupled from geographical proximity giving rise to *alternative* (“economic” and/or “institutional-relational”) *nonspatial proximities*.

Following the foregoing categorization, this chapter starts off by reviewing the archetypical a-spatial approach: that is, the linear model of innovation. The linear sequencing from basic into applied research and innovative products or processes leaves no conceptual room for geographical dynamics. The subsequent section looks at the literature that abandons the view of knowledge as a public good in order to explore the role of physical geographical distance in making knowledge a local quasi-public good. The fourth section places innovation “in context” by discussing (a) the influence of economic places – local agglomeration and specialization patterns – on the innovation process, by looking at how economists and geographers have tried to identify the type of sectoral specialization that is most conducive to innovation; (b) the role of local institutions is analyzed in the fourth section by reviewing the literature on regional systems of innovation (RSI) where the focus is on institutional-relational places. The fifth section will review recent research based on the multidimensional conceptualizations of proximity that broaden the analytical focus to nonspatial proximities as determinants of local innovative performance as also integrated approaches that combine and cross-fertilize the insights of various streams of literature. The final section concludes with some directions for future research.

34.2 The Linear Model of Innovation: The A-Spatial Benchmark

The linear model of innovation has for a long time been the most influential theoretical framework for the understanding of the economic impact of science and technology. It postulates that all innovations result from basic science

(Godin 2006): conducted in the research laboratories of universities and government research institutions, basic science produces new knowledge that is passed on to the applied science laboratories of private companies, where it is prepared for the translation into commercial products. The linear or “assembly-line model” (Ruttan 2001) conceives the innovation process as a one-way path:

Basic science → Applied science → Development → Commercialization and Diffusion

This view also implies that basic science creates positive externalities in the form of public knowledge: underinvestment in basic research must be expected in the absence of government intervention. The allocation of public resources to basic science is expected to maximize externalities that allow for the universal diffusion of knowledge as a public good.

Empirically, the reasoning behind the linear model lies at the core of econometric studies examining the link between R&D and patents, in the first instance, followed by that between patents and economic growth. These analyses are based on knowledge production functions (KPF), which allow for an investigation of the causal relation between productivity growth, unobservable knowledge capital, and its observable input (R&D) as well as output (typically patents), and further factors. Based on firm-level data, these studies are mostly conducted by “mainstream economists.”

The linear model of innovation has been particularly influential in the post-World War II when it shaped the US science and technology policies (Ruttan 2001) and remains popular with policy-makers in the twenty-first century, as evidenced by targets in terms of R&D spending to GDP ratios set in the EU’s Lisbon Agenda or by the contemporary policy focus on centers of excellence that still survives in the innovation policies of several countries. Two major reasons explain the lasting influence of the linear model. Firstly, the model conveys an unequivocal normative message: policy-makers should invest in basic research to maximize innovative potential. Second, national statistical offices and international organizations have reified “basic science,” “applied science,” and “development” into standard categories for the collection of data on innovative efforts, hardening the model as a concrete reference for policy discussions and transformed the linear model into a “social fact” (Godin 2006).

The most fundamental critique to this approach aims at the core of the model, that is, its linear character. The latter has been criticized for failing to reflect the complexity of innovation processes and the heterogeneity of its dynamics. These critics consider the production of new technological knowledge an interactive process between multiple agents. Since this process is assumed to involve continuous feedback, the advocates of this view reject the linear model’s conceptualization of innovation processes as a one-way sequence of steps. The creation of new knowledge is a socially embedded, interactive process. It is shaped by the interactions between innovative agents that, in their turn, are fundamentally influenced by physical space (that can facilitate or hamper their contacts) and by the places in which they are embedded being part of local industrial specialization processes, technological trajectories, and institutional modes of innovation.

34.3 Physical “Distance” Between Innovative Agents and Knowledge Flows

Once the view of knowledge as a pure public good – at the basis of the linear model – is replaced by a more realistic appreciation of its actual scope, geography as physical distance immediately becomes a fundamental component for the understanding of innovation processes. Knowledge has only a few of a public good’s characteristics: it is non-rivalrous and only to a limited extent excludable. In this regard, the literature on the role of geographical distance in innovation processes shares some common ground with the a-spatial linear model which assumes that knowledge production gives rise to external economies in the form of public knowledge. However, while in the linear model, the location of innovative agents is irrelevant to their capability to benefit from these externalities, the geographical literature considers knowledge as a spatially bounded quasi-public good whose circulation is largely restricted within the functional borders of the area where it is generated.

When looking at the spatial diffusion of knowledge flows, a crucial distinction is made between codified and tacit knowledge (Leamer and Storper 2001). The former is assumed to be relatively cheap to transfer since it can be expressed in a set of codes or instructions and distributed via communication channels (such as the Internet) and accessed by anybody familiar with the respective symbol system (e.g., language). Conversely, tacit knowledge is more expensive to transfer over long distances because – due to its higher complexity and context dependency – it is not codifiable (Leamer and Storper 2001). The relatively high cost of transferring tacit knowledge across space renders this type of knowledge geographically “sticky,” making face-to-face (F2F) contact an economically efficient means for its transmission. Encompassing verbal, physical, non-intentional, and intentional as well as contextual elements, F2F contacts allow for the communication of complex, contextual messages and minimize free rider problems by promoting the development of trust (Storper and Venables 2004).

The importance of F2F contacts can be interpreted as a pivotal factor underlying the spatial clustering of innovative activities: The complexity and context dependency of knowledge flows associated with innovative activities make the latter dependent on F2F – “an intrinsically spatial communication technology” (Rodríguez-Pose and Crescenzi 2008 p. 379). The dependency on F2F contacts may thus induce innovative actors to locate close to each other, which in turn leads to the emergence of geographical clusters of highly innovative agents.

In line with this conceptualization, geographical distance plays a major role in innovation processes: geographical proximity is deemed to facilitate the transmission of imperfectly appropriable but spatially sticky knowledge (Malecki 2010). Empirically, a large body of research on localized knowledge spillover (LKS) examines the importance of geographical proximity for the dissemination of knowledge (for a review, see Döring and Schnellenbach 2006): shifting from firm-based KPFs to regions as units of observation, this stream of literature finds

empirical support for the relevance of geographically mediated knowledge spillovers and identifies evidence of geographically bounded spillovers measuring their spatial extent (Döring and Schnellenbach 2006). A second stream of empirical literature has used patent citations to track the spatial diffusion of patented inventions, suggesting that patent citations display a high degree of spatial autocorrelation: inventors refer to previous patents originating in the same city more frequently than to a control group.

When it comes to the design of regional innovation policies, the consideration that geographical distance acts as a barrier for the diffusion of knowledge flows leads to the acknowledgment of geographical peripherality as a source of structural disadvantage. The emphasis on the spatial boundedness of knowledge flows may also be interpreted as warranting interventions aimed at minimizing the geographical distance between innovative actors in the public and private sector. Incubators and science parks are two examples of policy measures reflecting the idea that public policies can actively maximize spillovers promoting regional innovative output by providing infrastructure that allows for a spatial concentration of regional innovative activities.

However, “classic” studies on LKS are often based on indicators that capture the potential for spatially bound knowledge spillovers rather than actual flows/contacts between agents. The mechanisms underlying the transmission of knowledge spillovers remain underdeveloped, meaning that the concept of LKS is still largely a “black box” (Döring and Schnellenbach 2006): while some authors suggest that market transactions rather than externalities may explain local knowledge flows, others point out that members of epistemic communities may be connected by ties that transcend geographical proximity. The insufficient understanding of how knowledge is actually transferred between individuals located in the same geographical area impedes the formulation of a clear normative message to policy-makers.

In response to these criticisms, recent empirical work has focused more closely on the role of individuals as knowledge carriers and in particular on the mobility of knowledge-carrying workers and researchers. In addition, the literature has explicitly acknowledged that innovative agents cannot rely exclusively on local knowledge assets. Highly innovative actors benefit from a combination of “local buzz” (Storper and Venables 2004) – that is, the innovation enhancing local environment based on frequent F2F contacts of individuals who are colocated in a confined, typically urban place – and “global pipelines,” that is, communication channels formed by a differentiated set of “global” actors (different streams of literature have looked at multinational firms, diasporic communities, universities, and “star” scientists) that increasingly tap into pools of external knowledge bearing the associated communication cost/effort (Malecki 2010).

Only the most recent developments in economic geography theory (reviewed in Sect. 34.5) will overcome this dichotomous (local vs. global) conceptualization of knowledge transmission mechanisms developing more sophisticated frameworks of understanding.

34.4 Innovative Agents “in Context”: Local Specialization Patterns and Institutions

34.4.1 Economic Places: Industrial Specialization

Geographical distance between innovative agents is an important predictor of knowledge exchange costs. The communication of economically valuable potentially not codifiable/codified knowledge across large distances is possible but at increasing costs. However, a number of other characteristics of the local environment generate incentives for knowledge exchange and shape the synergies for innovation generation. In this context, a vast amount of literature has dwelt on the role played by specialization patterns by contrasting the innovation performance of both highly specialized and diversified economic environments that often coexist in both developing and developed countries.

A high degree of specialization facilitates the exchange of specialized, industry-specific knowledge. Occurring between firms active in the same industry, these Marshall-Arrow-Romer (MAR) knowledge spillovers are deemed to spur innovation. MAR spillovers are a typical feature of “classic” industrial districts. Conversely, “Jacobian spillovers” are associated with a diversified economic fabric, which is often found in big cities: the most valuable sources of knowledge of benefit to a firm lie outside its own industry. This view suggests that a diverse industrial structure allows for cross-industry knowledge flows that induce recombinant innovation.

The empirical literature suggests that both Jacobian and MAR externalities play an important part in enhancing innovation. Possibly due to differences regarding methodology and level of aggregation, analyses come to mixed, often conflicting results (for extensive reviews, see Beaudry and Schiffauerova 2009 and De Groot et al. 2009). Although part of the literature suggests that only specialization can be conducive to innovation, it must be stressed that MAR and Jacobian spillovers are not mutually exclusive (Beaudry and Schiffauerova 2009). Indeed, large cities can be simultaneously specialized in one or more sectors and simultaneously display a diverse range of further industries.

Specialization and diversification patterns have been harmonically combined into “economic places” by two sub-streams of literature. The first stream has combined specialization patterns with a product life-cycle perspective (Duranton and Puga 2001). Moving from a static to a dynamic view of the role of specialization patterns in the creation of new technological knowledge, innovation processes at different stages of the product life cycle rely on different types of knowledge spillovers. Firms develop new products in diversified urban contexts – termed “nursery cities” – benefiting from access to a greater variety of knowledge sources so that they can test new combinations until they identify the ideal production technology. Once production technology is standardized, firms relocate to specialized places as the focus shifts from radical to incremental innovations, and the ability to exchange knowledge with other firms from the same industry becomes more beneficial than having access to knowledge from a wide range of sectors.

In the nursery-city approach, both types of specialization patterns should coexist in a balanced system of cities, as they play different roles at different product life-cycle stages (Duranton and Puga 2001).

The second view that goes beyond the classic MAR versus Jacobian dichotomy proposes a more sophisticated understanding of sectoral diversity. The “related-variety” approach (Boschma et al. 2009) concentrates on cognitive proximity between sectors. Drawing on the notion of absorptive capacity, in a related variety framework, knowledge will necessarily “spill over” between any pair of industries: the identification and absorption of new knowledge requires a preexisting complementary knowledge. Related-variety industries share complementary competences (Boschma et al. 2009). Intermediate levels of cognitive proximity between related industries facilitate intersectoral knowledge flows conducive to innovation. Accordingly, neither specialization nor diversity per se enhances innovation: the former may lead to a too narrow knowledge base, whereas the latter might involve a lack of complementary knowledge across sectors. Instead, the composition of sectors in a region should ideally display an intermediate level of cognitive proximity between the different industries.

34.4.2 Relational-Institutional Places

While the industrial specialisation literature unquestionably abandons the a-spatial perspective of the linear model, it heavily concentrates on economic processes, essentially disregarding the institutional-relational dimension of territorial innovation processes. The concept of related variety does, however, share common roots with (regional) systems of innovation – the key components of institutional-relational places – and both streams are influenced by ideas from evolutionary economics and economic geography.

The systems of innovation (SI) perspective considers knowledge production as a nonlinear, interactive, and socially embedded process (Edquist 1997). SI literature adopts a systemic perspective and considers the creation of new knowledge as the result of evolutionary processes in complex systems. Its emphasis on multiple feedbacks between innovative agents sharply contrasts with the linear model's conceptualization of innovation as a one-way process. While in the linear model there are only three major types of innovative actors (corresponding to the categories of basic research, applied research, and product development), the SI approach allows for a great variety of participants in the innovative process. The organizations with which firms interact “to gain, develop and exchange various kinds of knowledge” (Edquist 1997 p. 1) include other enterprises but also government bodies, research institutes, universities, and banks (Edquist 1997). By embedding innovation in its social environment, this approach puts culture and institutions at the core of the analysis: habits, norms, and laws shape the relations between the innovative agents.

The literature has deployed the SI perspective in three major analytical perspectives: the sectoral, national, and regional levels. The sectoral systems of

innovation (Malerba 2006) highlight sector-specific patterns of knowledge production and suggest that the relative importance of different types of knowledge spillovers and learning varies across sectors. At the national level, different institutional settings and governance structures shape the synergies between innovative agents and their evolutionary trajectory. Combining the SI literature with concepts from economic geography that emphasize the local roots of innovation and learning, economic geographers and regional economists have extended the SI perspective to the regional level (Edquist 1997). The Regional Systems of Innovation (RSI) literature puts geography in the sense of institutional-relational places at the center of the analysis of spatial disparities in innovative performance. Iammarino (2005 p. 499) defines an RSI as “the localized network of actors and institutions in the public and private sectors whose activities and interactions generate, import, modify and diffuse new technologies within and outside the region.” From an RSI perspective, regionally specific modes of learning, technological trajectories, and knowledge bases constitute important reasons for regional disparities in innovative output.

The consideration of both “economic” and “institutional-relational” places has profound implications for innovation policies that depart from the “one-size-fits-all” approach supported by the “linear model.” The design of any innovation policy should reflect region-specific modes of knowledge production and industrial specialization patterns, making the in-depth understanding of the technological trajectory and existing knowledge base of each region the starting point for any innovation policy (Iammarino 2005; Asheim et al. 2011).

The RSI’s emphasis on interactive learning in “regionally embedded, institutionally supported, networks of actors” (Uyarra 2010 p. 125) implies that by simply increasing innovation inputs, policy-makers are unlikely to maximize a place’s innovative potential. The shift from individual actors to a systemic view calls for policy-makers to address the institutionally shaped relations between the components of the system. The rationale for public intervention comes from some kind of systemic failure, which calls for corrective measures aimed at improving the local institutional setup of a place. Cross-fertilizing the RSI perspective with the notion of related variety, Asheim et al. (2011) urge policy-makers to enhance innovation via “platform policies” facilitating knowledge flows between related sectors.

In comparison with the clear-cut normative message of the linear model, just how policy-makers should translate the RSI approach into practice is less straightforward. The approach has been criticized because it provides little guidance on instruments and measures appropriate for tackling systemic failures. The approach’s interpretative flexibility or “fuzziness” (Markusen 2003) renders its use more difficult for policy-makers. Equally, there are divergent views regarding the exact components and borders of an RSI. On the empirical side, a bias toward high-performing clusters has been also criticized (Uyarra 2010). A further weakness of empirical RSI studies stems from the lack of indicators appropriate to truly measure the performance of a system in terms of the quality of knowledge flows and interactive processes rather than in terms of absolute innovative output (Iammarino 2005).

34.5 Bringing Different Approaches Together: Nonspatial Proximities and “Integrated” Frameworks

As discussed in the previous sections, knowledge spillovers do not spread uniformly across space but exhibit strong distance-decay effects. While geographical proximity (*geography as physical distance*) facilitates the transmission of imperfectly appropriable but spatially sticky knowledge, the creation of new knowledge remains a socially embedded, interactive process. However, despite its potentially supportive role for the exchange of knowledge, geographical proximity constitutes “neither a necessary nor a sufficient condition” for learning processes (Boschma 2005 p. 62). Learning processes and communication are shaped by industrial specialization, technological trajectory, and institutional modes of innovation that are characteristic of specific *economic and/or relational places*. Consequently, the analysis of the geography of innovative processes calls for the joint analysis of the full set of physical, economic, and institutional conditions that make innovation possible. Economic geography theory has responded to this challenge in two ways. On the one hand, it has explicitly conceptualized the differential (and potentially independent) role of spatial and nonspatial conditions and, on the other, has fully explored the full set of their interactions.

In the first stream, Boschma (2005) has proposed a framework that introduces four nonspatial types of proximity, conceptually independent of physical distance: (i) cognitive proximity, referring to the degree to which agents share a common knowledge base; (ii) organizational proximity, defined as “the extent to which relations are shared in an organizational arrangement” (Boschma 2005, p. 65); (iii) social proximity, measuring social embeddedness based on friendship, experience, and kinship of relations between agents; and (iv) institutional proximity, which is based on agents sharing the same institutional rules and cultural habits. In this framework, cognitive proximity is considered as the only form of proximity that is a permanent prerequisite for interactive learning and innovation: without overlapping knowledge bases, learning is impossible – even if there is high geographical proximity between the agents. In this context, colocation and physical proximity may still play an important role on a temporary basis to establish contacts that are then maintained through the continuous presence of organizational, social, or institutional proximity. The positive effect of geographical proximity (*geography as distance* in our framework) might be more indirect and subtle than frequently assumed: it may help innovative actors to find the “optimal” balance between different a-spatial forms of proximity shaping “economic” and “institutional” places conducive to innovation.

Acknowledging that proximity can be defined independently of physical-metric considerations prepares the stage for an integrated view of the forces influencing regional innovation processes. The introduction of alternative proximities makes it possible to adopt a new perspective on the role of *geography as distance*. Nonspatial proximities provide the justification for knowledge flows in networks as described by Breschi and Lissoni (2005). Regions may thus use alternative proximities to overcome geographical distance and tap into remote knowledge

pools via global pipelines. Although this relativizes the significance of colocation, it is important to emphasize that Boschma's (2005) framework is nonetheless compatible with the concept of local buzz (Storper and Venables 2004): we may conceive local buzz as "cognitive, organizational, social and institutional proximity brought together in a reduced geographical environment" (Rodríguez-Pose and Crescenzi 2008 p. 383). From this point of view, alternative proximities influence both interregional and intra-regional knowledge flows. With respect, instead, to *economic places*, the notion of cognitive proximity is particularly fruitful for analyses of opportunities of learning across industries. As stressed in the related-variety perspective, cross-sectoral knowledge flows hinge upon the right level of cognitive proximity. As far as *institutional-relational places* are concerned, the idea that place-specific innovation systems display idiosyncratic modes of learning suggests that a lack of local institutional proximity may impede successful learning.

The second stream of literature focused more directly on the interaction between geography as economic places, institutional-relational places, and physical-metrical distance – while simultaneously acknowledging the importance of alternative, nonspatial proximities. Following an "integrated approach," any analysis of a region's innovative performance has to take five keystones into account: (i) the link between local innovative efforts and knowledge generation as typically emphasized by a-spatial approaches, (ii) the geographical diffusion of knowledge spillovers and the region's industrial specialization (representing geography as distance and geography as an economic place), (iii) the presence of networks based on alternative, nonspatial proximities, (iv) the genesis and structure of local and regional policies, and (v) the existence and efficiency of regional innovation systems, with the last two keystones reflecting geography as institutional-relational places (Crescenzi and Rodriguez-Pose 2011). The interaction of these five pillars shapes the creation of new knowledge in a region. In accordance with recent changes in economic geography theory, the importance of a-spatial networks and mobile capital with respect to global knowledge flows is underlined in this framework: the ability of local actors to establish external relations based on alternative proximities is assumed to determine the position of the region in global networks (e.g., where MNEs "pump" global knowledge into the local economy and "channel" the results of local innovative activities into global knowledge pipelines).

A number of subsequent empirical studies have built upon such integrated perspectives, aiming to shed light on the relevance of two or more of their elements. These contributions can be grouped according to their treatment of space/place on the basis of the categories developed in this chapter.

Table 34.1 provides an overview of the factors taken into account by recent contributions that in different ways contrast, compare, and/or interact with alternative conceptualizations of "geography and space." The columns of the table correspond to the four categories developed in this chapter: geographical distance (covered by studies that examine distance-based aspects such as LKS and agglomeration economies), "economic places" (regional sectoral specialization patterns), "institutional-relational places" (regional systems of innovation and other local socio-institutional conditions), and "alternative nonspatial proximities."

Table 34.1 Classification of recent developments in economic geography theory and their contribution to the understanding of territorial innovation dynamics

Authors (year)	Physical Distance	Economic places	Institutional-relational places	Alternative nonspatial proximities		
	Localized knowledge flows	Specialization	Regional systems of innovation	Institutional	Cognitive	Organizational
<i>Conceptualization</i>						
Boschma (2005)	X			X	X	X
Crescenzi and Rodriguez-Pose (2011)	X	X	X			X
<i>Empirical testing</i>						
Autant-Bernard and LeSage (2011)	X		X			
Crescenzi et al. (2007)	X	X	X			
Maggioni et al. (2007)	X			X		
Ponds et al. (2010)	X				X	
D'Este et al. (2012)	X				X	
Breschi and Lenzi (2012)	X	X				
Marrocu et al. (2013)	X			X	X	X

The first two rows of the table highlight the conceptual basis of the proposed classification in relation to the two conceptual papers reviewed above: Boschma (2005) for the conceptualization of nonspatial proximities and Crescenzi and Rodriguez-Pose (2011) for the “integrated framework” and the interaction between various geographical innovation dimensions. The second section of the table refers to “representative” empirical works that explicitly test the differential role of the various geographical aspects. The key “benchmark” and point of departure of all these papers is “geographical distance” whose impact on innovative performance is compared and contrasted with other relevant dimensions/factors. Autant-Bernard and LeSage (2011) look at “geographical distance” and “economic places” (in a sectoral perspective) by examining the impact of Marshallian and Jacobian spillovers both within and between regions by means of a knowledge production function approach. Their results shed light on the differential spatial extent of different typologies of knowledge flows suggesting that Jacobian externalities tend to decay more rapidly with geographical distance. In their comparison of the territorial dynamics of innovation in the USA and in Europe, Crescenzi et al. (2007) assess the influence of physical-metric, economic, and institutional-relational dimensions of geography. They use a modified KPF framework to account for intra-regional and interregional knowledge spillovers, sectoral specialization, and regional innovation systems conditions. Their study finds that the geographical processes governing knowledge production differ between Europe and the USA. While institutional-relational factors (in the form of social filters) are fundamental in both continents, the role of spatial distance differs substantially. In the USA, innovation is generated in relatively self-contained and more specialized geographical areas, while European regions rely heavily on the capacity to assimilate interregional knowledge spillovers. The importance of a-spatial networks and proximities is also acknowledged, in particular as far as the USA is concerned, although this perspective is not directly tested in the paper.

The influence of nonspatial proximities is directly examined by Maggioni et al. (2007) who compare the role of geographical distance against the influence of social proximity between research staff by looking at co-patenting data and EU-funded research collaborations by means of KPF and gravity models. They find that spatial proximity is of greater relevance to knowledge production than social proximity. Additional empirical work examining the relation between geographical distance and nonspatial proximities comes from the literature on university-industry collaboration. Again in a KPF framework, Ponds et al. (2010) examine the relative importance of geographical and social proximity (proxied by co-publication patterns), for the impact of academic research on regional innovation: social proximity makes it possible for knowledge spillovers to diffuse over large distances, suggesting that geographical proximity is of limited relevance for spillovers resulting from research collaboration. Opting for a different methodology, D’Este et al. (2012) employ a case-control approach for the examination of the role of geographical and organizational proximity in the formation of university-industry partnerships. They suggest that British companies in spatially dense clusters of technology-intensive industries establish connections with universities

largely independently of the university's location, whereas firms outside dense clusters seem to place more weight on geographical proximity when establishing their links with universities.

The work by Breschi and Lenzi (2012) points in a similar direction. They look at the internal and external network structures of US cities by linking the interactions of innovative agents at the microlevel with innovative output at the city level. They include social network indicators in a KPF in order to compare the innovation impact of the internal city-level coinvention network with the embeddedness of local inventors in global coinvention networks after controlling for the role of specialization patterns. The empirical results suggest that external linkages are only likely to improve regional innovative performance if they are combined with an appropriate intra-regional network structure that facilitates knowledge diffusion.

In a comprehensive attempt to disentangle the role of different forms of proximity, Marrocu et al. (2013) use an augmented KPF to investigate the relevance of the five a-spatial proximities proposed by Boschma (2005) and interregional spillovers. Coinventorship serves as a proxy for social proximity, while a similarity index based on the sectoral distribution of patenting activity in each pair of regions defines cognitive proximity. Organizational proximity is measured by the affiliation of applicants and inventors to the same organization, whereas country dummies are used to account for institutional proximity. The authors thus succeed in linking a-spatial networks based on alternative proximities at the individual level with innovative performance at the regional level. Their results suggest that cognitive proximity is always relevant, while geographical proximity is not the most important type of proximity for innovation processes, while the role of social and organizational proximity appears to be marginal.

This highly dynamic but still embryonic stream of literature, which explicitly aims at disentangling the innovative impact of various spatial and nonspatial factors, has not yet reached a consensus on the relative importance of different forms of proximity. The heterogeneity of the results is likely to stem from both methodological and operational differences. The estimation of knowledge production functions "augmented" in order to account for the impact of various proximities, although now customary in this literature, remains problematic due to the strong collinearity among the various proximities (whose impact the foregoing functions set out to isolate and compare) and the potential simultaneity between innovative performance and the evolution of nonspatial proximity relations. In addition, the use of patent data to measure both "proximities" and performance might generate additional measurement problems. Thus, in order to further advance our understanding of the transmission mechanisms underlying the geography of innovation, the KPF approach should be supplemented by other techniques able to directly model the formation of links and networks and their spatiality before assessing their impact on "aggregate" performance.

34.6 Conclusions

The conceptualization of geography in innovation literature has changed substantially since the heydays of the linear model. Persistent disparities in innovative

performance across space have motivated researchers to develop progressively a more sophisticated analysis of the role of space and places in innovation processes. From the physical-metrical approach of *geography as distance* to the emphasis on specialization and diversification patterns (*geography as economic place*) and *institutional-relational factors*, economic geography theory has substantially evolved in terms of its contribution to the understanding of technological dynamics.

While the abandonment of the linear model has always been at the very center of the geographical analysis of innovation processes, the most recent developments in the discipline have questioned the excessive emphasis on spatially localized processes that have long dominated the geographical approach. Geographical proximity has progressively lost its role as the single most important type of proximity to influence innovation processes. Cognitive proximity has emerged as a permanent requirement for interactive learning while social, organizational, and institutional proximity may act as temporary substitutes for geographical proximity. Geographical proximity remains to strengthen nonspatial proximities and helps innovative actors to find the right balance of nonspatial forms of proximity. The analysis of the systematic interactions among these different dimensions calls for progressively more “integrated frameworks” in order to understand territorial innovation dynamics.

These shifts in economic geography theory have important implications for innovation policies. The conceptualization of innovation as an interactive process occurring within complex innovation systems requires that policy-makers tackle linkages between actors rather merely making investments in basic research. Innovation policy starts from a profound understanding of a region’s idiosyncratic institutional setup, technological trajectory, and knowledge base. However, the identification of the potential barriers to innovative performance cannot be limited to the local dimension: understanding the region as the intersection of global and local knowledge flows implies that cooperation and networking should also be encouraged with remote partners in other regions and countries. At the same time, results indicating that academic spillovers can be mediated over longer distances via nonspatial proximities suggest that policy measures aimed at stimulating knowledge flows should not merely concentrate on the local level but rather adopt a national or even international perspective (Ponds et al. 2010). In addition, the acknowledgement of the crucial role of people as carriers of knowledge also implies that the generation and attraction of highly skilled individuals should be part of regional innovation policy (Tripll and Maier 2011; Marrocú et al. 2011).

Influential reports by the World Bank (World Bank 2009), the European Commission (Barca 2009), the OECD (2009), and the Corporación Andina de Fomento (2010) in different ways reflect recent theoretical changes in economic geography. While the World Development Report 2009 has the important merit of fully incorporating geography as distance and economic places into the formulation of development policies, the policy conclusions formulated by the OECD, the Barca Report, and the Corporación Andina de Fomento fully endorse an integrated territorial approach to innovation which takes full account of the role played by institutional-relational factors and nonspatial proximities.

The development of the economic geography theory of innovation has contributed toward a progressive shift in the policy paradigm from a purely “science and

technology” approach to the emphasis on agglomeration and spatial proximity that has characterized innovation policies targeting cluster development and firm incubators. However, the most recent evolution in the territorial theory of innovation opens the way to more balanced integrated policies that systematically account for the multifaceted influence of geography on innovation processes.

If the effectiveness of innovation policies can substantially benefit from the evolution of economic geography theory, a number of relevant aspects remain to be further explored both conceptually and empirically. From the conceptual point of view, further research is needed on the linkages between the microlevel of the individual innovative actors, the meso-level of their territorial interactions, and the diffusion channels of “macro” global flows of skills and knowledge. A sound theory for this complex set of processes is a necessary condition to open the “black box” of knowledge generation and diffusion. If the (increasing) importance of nonspatial proximities is now fully acknowledged, further work is needed on the reasons and the mechanisms that govern the development and the evolution of such proximities. In the same way as location theory aims to explain the colocation decisions on economic agents in physical space, it is necessary to explore the fundamental mechanisms that drive the development of nonspatial proximities between innovative agents in the cognitive space.

Conversely, the empirical analyses of the geography of innovation need to substantially broaden their scope both in terms of methodologies and use of available data in order to cope with increasing theoretical sophistication and new policy challenges (in both developed and emerging countries). If regional “aggregate” knowledge production functions have greatly contributed to the development of this field of research, it is crucial to reinforce microlevel analyses that can clearly target relevant actors and their behavior. Substantial progress is needed for a more detailed identification of the role of spatial and nonspatial networks in this context. In addition, the reliance on patent data has also led to the under-examination of non-patented forms of innovation including process and organizational innovation. The integrated use of different data sources (including firm-level innovation surveys such as the Community Innovation Survey) is certainly an important development in this direction, but the emergence of new and more sophisticated research questions calls for the collection of more sophisticated micro-data on the innovation and relational behavior of firms, individuals, and institutions.

Acknowledgments The author would like to thank Alexander Jaax for his excellent research assistance. Financial support by ESPON 2013-KIT Project is gratefully acknowledged. The author is also grateful to Andrés Rodríguez-Pose and Manfred Fisher for comments on earlier drafts of this chapter. The author remains solely responsible for any errors contained in this chapter.

References

- Asheim BT, Boschma R, Cooke P (2011) Constructing regional advantage: platform policies based on related variety and differentiated knowledge bases. *Reg Stud* 45(7):893–904
- Autant-Bernard C, LeSage JP (2011) Quantifying knowledge spillovers using spatial econometric models. *J Reg Sci* 51(3):471–496
- Barca F (2009) An agenda for a reformed cohesion policy. European Commission, Brussels

- Beaudry C, Schiffauerova A (2009) Who's right, Marshall or Jacobs? The localization versus urbanization debate. *Res Policy* 38(2):318–337
- Boschma RA (2005) Proximity and innovation: a critical assessment. *Reg Stud* 39(1):61–74
- Boschma RA, Eriksson R, Lindgren U (2009) How does labour mobility affect the performance of plants? The importance of relatedness and geographical proximity. *J Econ Geography* 9(2):169–190
- Breschi S, Lenzi C (2012) Net city: how co-invention networks shape inventive productivity in US cities. Working Paper, Università L. Bocconi
- Breschi S, Lissoni F (2005) Cross-firm inventors and social networks: localised knowledge spillovers revisited. *Ann Econ Stat* 79/80:1–29
- CAF (2010) Desarrollo local: hacia un nuevo protagonismo de las ciudades y regiones. Corporación Andina de Fomento, Caracas
- Crescenzi R, Rodriguez-Pose A (2011) Innovation and regional growth in the European union. Springer, Berlin/Heidelberg/New York
- Crescenzi R, Rodriguez-Pose A, Storper M (2007) The territorial dynamics of innovation: a Europe–United States comparative analysis. *J Econ Geography* 7(6):673–709
- D'Este P, Guy F, Iammarino S (2012) Shaping the formation of university–industry research collaborations: what type of proximity does really matter? *J Econ Geograph* (in press) doi:10.1093/jeg/lbs010
- De Groot HLF, Poot J, Smit MJ (2009) Agglomeration externalities, innovation and regional growth: theoretical perspectives and meta-analysis. In: Capello R, Nijkamp P (eds) *Handbook of regional growth and development theories*. Edward Elgar, Northampton, pp 256–281
- Döring T, Schnellenbach J (2006) What do we know about geographical knowledge spillovers and regional growth?: a survey of the literature. *Reg Stud* 40(3):375–395
- Duranton G, Puga D (2001) Nursery cities: urban diversity, process innovation, and the life cycle of products. *Am Econ Rev* 91(5):1454–1477
- Edquist C (1997) Introduction. In: Edquist C (ed) *Systems of innovation: technologies, institutions, and organizations*. Pinter, London, pp 1–35
- Godin B (2006) The history of the linear model of innovation: the historical construction of an analytical framework. *Sci Technol Human Values* 31(6):639–667
- Iammarino S (2005) An evolutionary integrated view of regional systems of innovation: concepts, measures and historical perspectives. *Euro Plan Stud* 13(4):497–519
- Leamer EE, Storper M (2001) The economic geography of the Internet age. *J Int Bus Stud* 32(4):641–665
- Maggioni MA, Noselli M, Uberti TE (2007) Space versus networks in the geography of innovation: a European analysis. *Papers Reg Sci* 86(3):471–493
- Malecki EJ (2010) Everywhere? The geography of knowledge. *J Reg Sci* 50(1):493–513
- Malerba F (2006) Sectoral systems: how and why innovation differs across sectors. In: Nelson R, Mowery DC, Fagerberg J (eds) *The Oxford handbook of innovation*. Oxford University Press, Oxford/New York, pp 380–406
- Markusen A (2003) Fuzzy concepts, scanty evidence, policy distance: the case for rigour and policy relevance in critical regional studies. *Reg Stud* 36(6/7):701–717
- Marrocu E, Paci R, Usai S (2013) Proximity, networking and knowledge production in Europe: what lessons for innovation policy? *Technological Forecasting and Social Change* (in press)
- OECD (2009) *How regions grow*. OECD, Paris
- Ponds R, van Oort F, Frenken K (2010) Innovation, spillovers and university–industry collaboration: an extended knowledge production function approach. *J Econ Geography* 10(2):231–255
- Rodríguez-Pose A, Crescenzi R (2008) Mountains in a flat world: why proximity still matters for the location of economic activity. *Cam J Reg Econ Soc* 1(3):371–388
- Ruttan VW (ed) (2001) *Technology, growth, and development. An induced innovation perspective*. Oxford University Press, New York
- Storper M, Venables AJ (2004) Buzz: face-to-face contact and the urban economy. *J Econ Geography* 4(4):351–370

- Trippl M, Maier G (2011) Knowledge spillover agents and regional development. In: Nijkamp P, Siedschlag I (eds) Innovation, growth and competitiveness. Springer, Berlin/Heidelberg/New York, pp 91–111
- Uyarra E (2010) What is evolutionary about “regional systems of innovation”? Implications for regional policy. *J Evol Econ* 20(1):115–137
- World Bank (2009) World development report reshaping economic geography. World Bank, Washington, DC

Henry G. Overman

Contents

35.1 Introduction	668
35.2 Empirical Analysis: Data	670
35.3 Empirical Analysis: Causality	672
35.4 Policy Evaluation	675
35.5 Conclusions	680
References	681

Abstract

This chapter is concerned with the process by which geographical economics influences policy. It considers a number of barriers that limit this influence focusing specifically on the availability of data, the limitations of spatial analysis, and the role of the evaluation of government policy. It considers why these problems present such significant barriers and proposes some solutions. In terms of the availability of data, the chapter explains why problems concerning the correct unit of analysis and measurement error may be particularly acute for spatial data (especially at smaller spatial scales). Resulting concerns about the representativeness of data and the mismatch between functional and administrative units may further hamper interaction with policy makers. For spatial analysis, the major problem concerns the extent to which empirical work identifies the causal factors driving spatial economic phenomena. It is suggested that greater focus on evaluating the impact of policies may provide one solution to this general identification problem.

H.G. Overman

Spatial Economics Research Centre and Department of Geography and Environment, London
School of Economics and Political Science, London, UK
e-mail: h.g.overman@lse.ac.uk

35.1 Introduction

In most countries, economic prosperity is very unevenly distributed across space. Regions, cities, and neighborhoods seem to be very unequal. This is true if we look at average earnings, employment, education, and almost any other socioeconomic outcome. Regional policy, urban policy, and even neighborhood policy are all largely based on concerns about these kinds of disparities, and tackling these persistent disparities is a key policy objective in many countries. Providing a rigorous understanding of the nature, extent, causes, and consequences of these disparities has been a key motivation behind the development of geographical economics (broadly defined).

This chapter focuses on the policy response to these disparities and specifically on the interaction between academic research and spatial economic policy. In the limited space available, it is clearly not possible to summarize all the available research that has relevance to policy makers concerned with spatial economic policy. Instead, this chapter considers the process by which research informs policy, focusing specifically on the role of empirical analysis. In doing so, the chapter considers criticisms of existing empirical work, provides an introduction to means of evaluating the impact of spatial policies, discusses the major barriers to interaction, and makes some suggestions on how these might be addressed in future research. The last two of these issues have received some consideration by Markusen (2003) and Martin and Sunley (2011) from the perspective of economic geography “proper.” In contrast, this chapter is specifically concerned with geographical economics (i.e., the research field that has evolved at the interface between economics and geography and which this chapter treats as synonymous with spatial economics). However, it is clear that many of the issues apply more generally in terms of the impact of research on policy making.

The chapter focuses specifically on the role of *empirical* analysis and policy evaluation in informing policy. The strong theoretical bias of the new economic geography (Krugman 1991) means that many of the issues concerning the application of theory to policy have received fairly detailed consideration in the literature (see, e.g., Baldwin et al. 2003; Combes et al. 2005). Duranton (2011) provides a diagrammatic framework which carefully outlines many of the central issues. This literature reaches two broad conclusions. First, from a positive perspective (i.e., what will be the impact of a specific policy), the theoretical literature is better placed to provide general guiding principles rather than detailed answers. Second, our theoretical understanding of what policy should do (i.e., normative analysis) is far less developed and, as usual, depends crucially on assumptions about the relevant objective function. In short, from an academic perspective, while theoretical analysis is not always sufficiently well developed to be useful in guiding policy, the problems are at least well understood.

From a policy makers’ perspective, these theoretical issues are arguably second order. Instead, the fundamental concern is whether or not stylized formal modeling can ever provide “real-world” insights. Of course, these concerns are not unique to

policy audiences nor to spatial policy. However, assuming some policy makers do not hold such reservations on the validity of formal modeling (or, at least, are willing to set them aside), the central issue from a policy perspective becomes the provision of empirical evidence about the applicability of the underlying theory. This issue and the barriers faced in providing such evidence in the specific context of policy formation have received far less attention in the literature. It is for this reason that this chapter focuses on the role of the empirical analysis of spatial data in policy making.

Of course, questions concerning the empirical analysis of spatial data go beyond the role this might play in assessing the validity of formal modeling. Indeed, for most policy makers, this would be a distinctly second-order concern. Instead, experience suggests that policy makers look to empirical analysis to do (at least) three things: describe the problem, assess the underlying causes, and evaluate the alternative policy responses. In fact, these three roles are not so far removed from how many geographical economists would prefer to structure empirical research.

In an ideal world, theoretical modeling would deliver predictions about the underlying causes of spatial disparities. Appropriate data would then be used to describe these disparities and to test the validity of predictions from the theory. Assuming that the data support the predictions from the theoretical model, one could then use the model to think through the impact of alternative policy responses that have not yet been implemented. More recently, it has also been recognized that this logic can be reversed, with theory providing predictions about the impact of policy and empirical evaluation of policy that has been implemented then used to test the underlying theory. That is, assessing the causal impact of existing policies may be useful in increasing our theoretical understanding of how the spatial economy works, what causes spatial disparities, and what, if anything, policy might do to address these disparities. In addition, there is considerable interest in establishing the causal impact of existing policy independent of what it can tell us about theory.

Geographical economics faces a number of barriers in addressing each of these questions. Data availability can hamper the provision of basic descriptive statistics, as well as further empirical analysis. Spatial research (by academics) often fails to pay enough attention to the central question of identifying the causal mechanisms at work. Finally, and related, much policy evaluation (by governments and consultants) fails to identify the causal impact of policies, often despite claims to the contrary. On all three dimensions, and particularly with respect to policy evaluation, this chapter will argue that the empirical literature analyzing spatial data often falls somewhat short of the standards set by other fields of economics. This partly reflects the inherent difficulty of spatial analysis but also stems from a failure by some researchers to adopt methodological developments that might improve analysis. The empirical literature is only beginning to address this shortcoming.

The rest of this chapter is structured as follows. The next section focuses on the availability of suitable data, the starting point for better empirical analysis. The two subsequent sections deal with questions of causality and the policy evaluation of spatial policies, while a final section briefly concludes.

35.2 Empirical Analysis: Data

Problems of data availability depend on the policy context, but the most common problem tends to arise from the lack of available data at the appropriate spatial scale. Even at large spatial scales, these problems can be acute. For example, in the UK, even basic statistics to describe spatial disparities across cities are not easily available (DCLG 2006). In other countries, while data for larger spatial units is readily available (e.g., for US metropolitan areas), more detailed data (e.g., on firm location) may not be available or may be subject to quite restrictive access arrangements. From a policy perspective, the lack of spatial data that can be used to generate statistics to describe the problem at hand represents a major barrier to the use of geographical economics in policy making. This barrier is arguably greater for geographical economics than for some other disciplines similarly concerned with spatial disparities but which rely less on formal modeling and quantitative empirical analysis.

Even when data is available, however, there remains the fundamental problem that for most issues in geographical economics, the correct unit of analysis is difficult to define. For example, researchers in the field of international economics can often use nation-state boundaries to define the appropriate unit of analysis because these boundaries generate significant barriers to factor mobility (and differences in factor availability underpin many theories of international trade). For the spatial researcher, in contrast, city or regional boundaries are often no more than administrative creations. For some problems, where administrative units form the appropriate unit of observation, such data might be sufficient. This may partially explain, for example, why local tax competition and public good provision have been so extensively studied in the empirical literature. For other outcomes, such as economic growth, administrative units may provide a very poor substitute for properly defined functional economic areas (see Cheshire and Magrini (2009) for further discussion).

Problems concerning the definition of a suitable unit of analysis and the availability of data for these units often become more serious at smaller spatial scales. For example, the appropriate definition of a “neighborhood” is a major concern for researchers interested in identifying the importance of neighborhood effects (i.e., whether neighborhood composition affects individual outcomes over and above any effect of individual characteristics). As was the case for larger spatial units, these definitional problems may represent a more serious disadvantage for geographical economics than for disciplines that adopt qualitative approaches to consider the existence of neighborhood or peer effects. An ethnographic study, for example, can easily accommodate self-defined notions of neighborhoods. In contrast, (spatial) econometric analysis requires neighborhoods to be formally defined so that data can be collected that characterizes the structure of the neighborhood. Nor is this definitional problem the only, or even the most significant, barrier to econometric analysis in this area (and many related ones concerned with feedback between units of analysis in the outcome of interest). We return to this issue below.

Even when data is available for something approximating the correct unit of analysis for the question at hand, there may be considerable measurement error present. Of course, measurement error is usually present in nonspatial data, but the problem is more pronounced for spatial data because all of the standard problems occur (e.g., is employment correctly defined, measured, and recorded), but there is an additional spatial “allocation” problem. This spatial allocation problem arises because the construction of spatial data for any specific unit of analysis requires dots on a map to be allocated to units in a box (see Duranton and Overman (2005) for further discussion). Inaccuracies can occur both in terms of the geographical location of the “dot” (e.g., the spatial coordinates of zip codes) and the boundaries of the box. Uncertainty over the correct definition of the unit of analysis, as discussed above, exacerbates these difficulties.

These measurement problems become more profound as the spatial resolution of the data increases because any absolute measurement error translates into greater relative error at smaller scales. Even if points are accurately allocated to geographical units, researchers using sampled data face an additional problem at small scales: For a given sampling frame, smaller spatial scales reduce the average sample size in any given geographical unit. As discussed in Duranton and Overman (2005), these problems can sometimes be avoided by working in continuous space. Whether this is a solution depends, however, on the problem at hand. When the issue is one of individual behavior (e.g., whether individual labor market outcomes are affected by employment accessibility), it may make sense to work in continuous space using geo-coded individual data. In contrast, when the interest is in broad spatial patterns (e.g., what causes differences in city growth), analysis in continuous space, based on individual geo-coded data, may not be helpful.

Regardless of whether or not switching to continuous space may help with analysis, it often will not solve the problems that poor data create in terms of generating descriptive statistics for spatial units of observation. Experience suggests that a lack of suitable data and the resulting inadequacy of *descriptive* statistics represent a significant barrier in using geographical economics to inform spatial policy making. As argued in Overman (2010), the increased use of geographical information systems (GIS) is slowly helping to solve many of these problems of data availability. GIS are helping reduce measurement error as well as making more data available by facilitating the reconciliation of data for different non-nested spatial units. The increased availability of geo-referenced data also allows researchers increasingly to avoid the need for arbitrary discretizations of data (because they allow the researcher to construct data for appropriate spatial units). Finally, new types of data are helping increase our understanding of spatial economic phenomena.

Interestingly, however, even if data availability becomes less of an issue in terms of analysis, a lack of descriptive statistics for specific administrative units may continue to cause a problem in terms of the interaction with policy makers. A major part of the problem stems from the fact that while these administrative units may be arbitrary from an analytical perspective, they are hugely important from a policy makers’ perspective. Policy makers want to know how these administrative units

are performing partly as an input into decision making but also because their performance is often assessed by comparison to other similar units. As a result, even when such data might not be a particularly useful guide for relative performance (if, say, a labor market boundary spreads beyond the administrative boundary), it will still be of great interest to policy makers.

The problem is further compounded when it comes to the empirical analysis of the underlying causes of spatial disparities. Specifically, in the absence of descriptive statistics based on representative data for particular places, many policy makers think that no progress is possible. This problem appears to arise because many policy makers struggle with the idea that sampled data can be informative about spatial processes unless the data is representative (which they equate with the production of “accurate” descriptive statistics for specific places).

Of course, this problem of representativeness is not unique to spatial settings, but experience suggests that it seems to be particularly important in terms of policy makers’ concerns about empirical analysis in this area. One possible underlying source of the problem may arise from the belief that each location is somehow unique either in terms of its characteristics or its responsiveness to policy initiatives (or both). However, as is increasingly recognized in applied microeconomics, such heterogeneity (including in the response to changes) may require care to be taken in interpreting statistical estimates but does not invalidate regression analysis of the problem at hand (see, e.g., the discussion of local average treatment effects in Angrist and Pischke (2009)). Given that these insights are often poorly understood by many researchers, it is no surprise that they have not had much influence to date on policy makers. It is not so clear why the criticism seems to have such bite in terms of the empirical analysis of spatial data. It would be interesting to know, for example, whether the problem is particularly acute because of the tendency of other disciplines’ strong emphasis on the uniqueness of location as an argument against quantitative approaches to spatial problems. Regardless of the underlying reason, the lack of large samples for administrative spatial units remains a barrier to informing policy making at all spatial scales even though it need not be.

35.3 Empirical Analysis: Causality

As discussed above, even when appropriate spatial data is available, the second broad set of problems concerns the type of empirical analysis that has been traditionally undertaken using such data. These problems are discussed in detail in Gibbons and Overman (2012) who argue that the biggest problem stems from the fact that traditional spatial econometric and statistical analysis has not paid sufficient consideration to the crucial issue of identification. This has profound implications for our ability to understand the causes of spatial disparities and for empirical analysis to influence the development of policy. To understand why, one needs to consider the way in which the empirical analysis of observed data might allow us to understand causality.

In many fields of economics, empirical research is increasingly concerned with questions about causality (Angrist and Pischke 2009). That is, questions of the type “if we change x what do we expect to happen to y .” This is particularly the case in fields focused on individual (microeconomic) rather than aggregate (macroeconomic) behavior. Although geographical economics is clearly concerned with both levels of analysis, if it is to be useful in policy making, then these types of questions must take center stage. After all, policy usually seeks to change some x in order, hopefully, to achieve some desired change in outcome y . Even when policy can directly influence the outcome of interest, we still need to understand how economic agents adjust to any change so that we can establish what will happen after this adjustment has taken place.

The fundamental challenge to answering these questions for (most) economic data is that the determinants (x) are not randomly assigned. This is certainly the case for many policy interventions, when x (e.g., investment in the transport network) will often be specifically set to (partially) reflect differences in the outcome of interest y (e.g., the level of GDP). As a result, in real-world data, we jointly observe x and y , so we lack the counterfactual. That is, what would have happened if x had been set at some different level? This is a problem because it is the comparison of actual outcomes to this counterfactual that identifies the causal impact of determinant x on outcome y . Fortunately, applied economics has come a long way in its efforts to find credible and creative ways to answer such questions by constructing counterfactuals from observational data. Unfortunately, however, such methods have not been widely applied in much applied analysis of spatial data, particularly in analysis undertaken using the standard spatial econometrics toolbox.

Instead, much applied spatial econometrics research assumes that we know the way in which spatial interactions occur, writes down the corresponding spatial econometric model, and estimates the parameters by nonlinear methods such as (quasi-)maximum likelihood. Questions of identification (i.e., does an estimated correlation imply that some determinant x causes outcome y ?) have generally been addressed by asking which spatial processes best fit the data. While this sounds straightforward, Gibbons and Overman (2012) explain that it is very hard to distinguish between alternative specifications that have very different implications for which causal relationships are at work. This fundamental identification problem, and the lack of attention given to it, significantly reduces the usefulness of this kind of spatial econometric analysis for policy making.

In practice, analytical capacity constraints limit the extent to which many government departments can engage with quantitative (econometric) analysis. Given this more general problem, it is perhaps no surprise that spatial econometric model specifications and estimation are sufficiently complex that research in this tradition has often proved very hard to communicate to policy makers. Coupled with concerns about the underlying secondary data, this can often lead policy makers to prioritize research which focuses on carefully describing the nature of spatial disparities rather than properly identifying the underlying causes. As discussed above, this tendency is reinforced by political interest in outcomes for specific administrative units. Again, it would be interesting to understand why this

tendency to conflate description with analysis is so pronounced in the area of spatial disparities and policy making.

If data availability and the type of analysis undertaken with available data represent significant barriers to informed policy making, a further barrier arises because, even when these problems are recognized, they can be very hard to address. The fundamental reason for this is that for many spatial economic phenomena of interest, suitable identification strategies can be hard to develop.

Researchers are making progress in this area, but, as argued in Gibbons and Overman (2012), progress is likely to remain slow unless issues of identification are given far more precedence in spatial empirical analysis. They suggest that these issues need to be put at center stage and discuss strategies for dealing with identification in spatial settings. One possibility is to use explicit sources of randomization that occur as a result of institutional rules and processes. For example, the random allocation of dorm-mates has been used by Sacerdote (2001) and others to study peer effects in college grades. Randomization often does not solve all identification problems, but it does reduce problems arising from the self-selection of individuals into groups. However, when it comes to areas of substantive policy interest, there are several barriers to randomization, especially in terms of exposure to policy interventions.

The major one is arguably political. While many academics are comfortable with randomization (e.g., because they are willing to start with the assumption that a policy will have no effect), this is a far harder proposition for policy makers. For example, if a policy maker starts with the assumption that policy will be beneficial, then randomization generates ethical concerns that those most in need might not be treated. These ethical, and other practical, issues have been extensively discussed by many of the so-called randomistas who advocate the use of random trials in the development context (see, e.g., Banerjee and Duflo 2009). While they are clearly making some mileage in specific circumstances, the general arguments are not yet won, even in circumstances in which randomization would be exceptionally helpful.

In addition to this central problem, large-scale field experiments such as the Moving to Opportunity program are rare and costly (and still suffer from very difficult to avoid design flaws). On the other hand, small-scale experiments suffer from concerns about external validity (i.e., the extent to which the results would generalize to other contexts). Such concerns about external validity, although not expressed in this way, may have particular bite for policy makers in the area of spatial policy who, as discussed above, often think of every place as being somehow unique. For all these reasons, it is hard to imagine policy makers agreeing to experiments to answer many spatial questions (even if such experiments could be designed), and it is therefore unlikely that randomization will represent a way forward for many areas of interest.

In the absence of suitably randomized data, appropriate instrumental variable strategies may represent an alternative way of circumventing the reverse causality problems that bedevil much spatial analysis (particularly in areas important to policy making). That said, it is often hard to think of suitable instrumental variables

in situations where we are interested in “area effects” that arise because of feedback to the outcome of interest from the outcomes of other economic agents located nearby. As has been understood for some time and formalized by Manski (1993), econometric analysis of such “endogenous social effects” faces severe identification challenges. This can make it very difficult to assess whether such effects are occurring or whether the appearance of interaction arises because of underlying similarities between nearby units of observation. It is interesting to note that empirical analysis looking at neighborhood effects (e.g., the impact of neighborhood on schooling) tends to find little evidence of strong effects when it carefully addresses these identification issues. This is in direct contrast to more qualitative approaches. As argued by Cheshire et al. (2008), this may partly explain why geographical economics has had relatively little influence on policy aimed at neighborhood “mixing” and other initiatives to “mitigate” neighborhood effects. It may be that the best we can do in these circumstances is to make policy makers realize the difficulties inherent in distinguishing these alternatives and point out that in research to date, the more careful the identification strategy, the less evidence there is of interaction through endogenous social effects.

In other spatial settings, however, where interest is not limited to endogenous social effects, it is increasingly possible to develop effective instrumental variable strategies as a result of policy designs, institutional rules, and natural environmental features (or even better, changes in these factors). Overman (2010) and Gibbons and Overman (2012) provide many concrete examples. One significant problem remains; however, these strategies can be very hard to explain to policy makers who do not fully understand the need for careful identification strategies (at least for those with little or no economics training).

One possible way to interest policy makers in these issues is through the use of identification strategies based on the details of existing policy interventions. Policy makers often need to evaluate the impact of policy. In addition, specific policy features may also help with the identification of causal factors at work in spatial processes. It would appear, then, that there may be two linked arguments for focusing greater efforts on credible policy evaluation. First, one would hope that effective policy evaluation should be a key input into policy development. Second, such policy evaluations may provide useful identification strategies to increase our understanding of the way the spatial economy works. Because of the possibilities this presents, it is worth considering these issues in some detail, and it is to this that we now turn.

35.4 Policy Evaluation

Policy specific outputs (e.g., the number of workers trained or firms assisted) are increasingly well monitored by governments. In contrast, many formal (i.e., government sponsored) evaluations of policies that seek to look at outcomes do not use credible identification strategies to assess the causal impact of policy interventions. As for spatial empirical analysis more generally, it could be argued that this

problem appears to be particularly acute for spatially targeted policies. Once again, these problems partly stem from the difficulty in coming up with identification strategies in spatial settings. That is, in assessing, what would have happened to the unit of analysis (the area, firm, worker, etc.) in the absence of the policy intervention? As emphasized by the literature on program treatment effects (see, e.g., DiNardo and Lee 2010), solving this problem requires the construction of a valid counterfactual that can then be compared to observed outcomes. In this section, we argue that, despite the difficulties, such an approach can be applied to many spatial policies and that the resulting evaluation can be informative about both the effect of policy and the spatial economic processes at work. Some concrete examples, mostly drawn from the USA and UK, will be used to help clarify the issues.

Let us start with the example of Enterprise Zones (also known as Empowerment Zones in the USA and referred to below as EZs). These spatially targeted policies aim to improve economic outcomes (e.g., employment and number of businesses) in deprived areas. To identify their causal effect, we need to figure out what would have happened in these areas in the absence of intervention. One possible identification strategy is to compare these areas to other similar areas that were not targeted by the policy. Actually, for many government-funded reports, even this simple strategy would substantially improve the quality of evaluations. From an academic perspective, however, such simple comparisons remain problematic because they require very strong identifying assumptions. Specifically, unless we have an exhaustive list of area characteristics that might influence local economic outcomes, we might worry that some unobserved characteristic of areas drives both the decision to target the area *and* outcomes in that area. In this case, we might wrongly attribute any change in outcomes to the policy when, in fact, it is driven by unobservable area characteristics. Much of the recent improvement in the evaluation of program treatment effects has come from novel ways of addressing this problem combined with a refined understanding of how to interpret the resulting estimates.

One possibility is to compare outcomes for those areas that receive funding to those areas that applied for, but did not receive, funding. This strategy has been used by Busso et al. (2010) in their recent evaluation of the US Empowerment Zone policy. Such a strategy can be highly effective in removing the influence of many unobservables that might bias estimates of policy impact, especially if restrictions to funding limit the number of areas treated so that selection among the applicants is less likely to be driven by these unobservable characteristics. More recently, the UK government announced that 29 sites will compete to host 10 new Enterprise Zones. As for US Empowerment Zones, with these new Enterprise Zones, the 19 sites that lose in the competition may provide a reasonable control group for the 10 that win. Comparing outcomes for the two groups will then tell us whether those that won the competition actually do better, and we may be willing to attribute this to the impact of the policy. Analysis could also compare those that entered the competition to areas that appear to be similar but that did not enter the competition (to see whether those that entered the competition somehow differ from those that do not). For these kind of strategies to achieve identification of the causal effect of the policy requires that, conditional on observable characteristics of areas, treatment is not correlated

with any unobservable characteristic that directly influences the outcome of interest.

The timing of policy interventions may provide another possible source of identification. For example, EZs given money early on should start improving before those given money later. If they do not, that raises questions about whether treatment caused any improvement (or decline) or instead whether this was caused by some other factor (such as changes in the macroeconomy). These strategies have received recent application in the literature including the work by Busso et al. (2010) on US Empowerment Zones. For this strategy, identification of the causal effect of the policy requires that, conditional on observable characteristics of areas, *timing* of treatment is not correlated with any unobservable characteristic that directly influences the outcome of interest.

Even in situations where the researcher cannot be sure that decisions to fund (or the timing of funding) are uncorrelated with all unobservable characteristics that directly influence the outcome of interest, we may believe that this condition holds for *marginal* decisions. Imagine, for example, that the government makes its funding decisions on the basis of a ranking of projects from best to worst. Such detailed assessment of projects often occurs after a rougher process has ruled out the weakest projects (so the sample of projects subject to the more detailed ranking may be those that make it through this first screening process). If a researcher has access to the ranking of projects, then this would allow the comparison of outcomes for otherwise similar areas that were just “above the bar” (and so got treated) to outcomes for areas just “below the bar” (who did not get treated). Sometimes, the criteria for treatment will be based on some observable characteristic of areas rather than some ranking based on the quality of bids submitted to the program under consideration. Then areas that just satisfy the criteria and so get treated can be compared to areas that just fail to satisfy the criteria and so do not get treated. Some policies, such as the UK’s Local Enterprise Growth Initiative may use a combination of cutoff criteria and competition to decide who gets treated (from among those that are eligible).

Applications of such regression discontinuity designs to spatial economic policies include Baum-Snow and Marion (2009) and Dachis et al. (2011). As discussed further in Lee and Lemieux (2010), these discontinuity designs can be used to identify the causal effect of policy, providing that applicants do not have full control over the characteristics that determine treatment. Notice that this is a weaker requirement than having no control (so that treatment need not be completely random across all areas) but comes at some cost in terms of the extent to which estimated effects generalize to areas that are further away from the policy cutoff. This is sometimes characterized as involving a trade-off between internal and external validity (i.e., the researcher gets good estimates of the causal effect for areas around the threshold, but it is not clear whether these would generalize to areas away from the threshold). This distinction provides one example of how the recent program treatment literature has clarified our understanding of how to interpret estimated parameters as well as how to estimate them in the first place.

So far there is nothing specifically spatial about these identification strategies (other than the fact that the policy intervention occurs in specific places) which have been more widely used in other applied microeconomic literatures (particularly in development, education, and labor). However, the fact that the policy intervention occurs in specific places and that these places have a geographical location provides a further source of discontinuity which may be useful in achieving identification. Specifically, we can use “spatial differencing” to compare treated areas to *nearby* non-treated areas. If unobservable characteristics vary smoothly over space, then such a comparison may help control for unobservable characteristics that affect both treatment and outcomes. As with regular (nonspatial) discontinuity designs, the validity and interpretation of the resulting parameter estimates depend crucially on how the borders of treated areas are determined and what happens to the unobservable characteristics of areas at those borders. If unobservable characteristics vary continuously at the border, then spatial differencing may give us the causal effect of the policy even if policy assignment is nonrandom (providing that policy makers do not have perfect control over the location of the boundary that determines the policy area). Even if unobservable characteristics do not vary continuously at the border, spatial differencing may still help if it eliminates larger spatial trends, making it easier to find suitable instruments for the spatially differenced variables (see Duranton et al. (2011) for further details and an application to the impact of local tax rates on employment).

A further complication arises when using spatial differencing if treatment effects spill over geographical boundaries to impact non-treated areas. This spillover might be positive (often referred to as a multiplier effect) or negative (often referred to as displacement). Regardless of the sign of the effect, if the interest is in the overall aggregate impact of the policy for an area that extends beyond the boundary of the treated zone (as it might be, e.g., for Enterprise Zones) such spillovers significantly complicate interpretation of estimated coefficients. Specifically, in the presence of positive spillovers, estimates of the effect of policy are biased downward and vice versa for negative spillovers. These issues are discussed further in Neumark and Kolk (2010), but the literature is only just beginning to grapple with the resulting complications.

Official evaluations of government policies (i.e., those paid for and sponsored by government) usually make little, if any, use of these program features to help identify the causal impact of policy. For a geographical economist, this significantly complicates the interaction with policy makers, because reports that are less careful about causality are often willing to make much broader claims about the impact of a policy (and how that impact was achieved). As a result, policy makers face a difficult trade-off when trying to decide how to evaluate policies. Wide-ranging “evaluations” that are less careful about causality *appear* to provide more information as an input in to the policy-making process. Taken at face value, such evaluations allow policy makers to both assess value for money and make changes to policy, while appearing to take into account evidence about the impact of the policy. In contrast, empirical research in the program treatment effects tradition often makes fairly narrow claims about whether the policy has a causal impact (and

then, sometimes, only for a particular part of the population depending on the methods used).

Of course, there are a number of arguments in favor of an approach which focuses on a narrower range of issues concerning the causal effects of policy. First, and most important, a policy evaluation that focuses on causal effects should substantially improve our understanding of whether policies such as Enterprise Zones have *any* net impact (including possibly whether or not they generate or mainly displace economic activity). This would help future governments when they decide whether to maintain or reintroduce such a scheme. In addition to this core reason, it is also interesting to note that in many circumstances, government could get this type of policy analysis at little cost because this kind of evaluation has the potential to be published in top academic journals (cf. a number of the references provided above). Such “open evaluation” will not work for all policies (because the degree of academic interest will usually depend on the extent to which the policy “design” allows causal effects to be identified), but it could work for a good proportion of them. In short, when appropriate, policy evaluation of this kind does not need to be big, expensive, and centralized. Instead, it can be outsourced by using open evaluation in the academic (and wider nongovernmental) community. A major barrier to such an approach to evaluation is, once again, the availability of data. But now the issue concerns the availability of information on the government policy to be evaluated. A first step in moving toward a more open evaluation model would require good information to be recorded at all stages of the policy-making process – for example, whether selection of projects is competitive, how decisions are made, what is the location and timing of intended and actual expenditure, and what types of expenditure (buildings, capital grants, training?) are funded. Information on bids needs to be available whether successful or not. Nearly all of this information will be available and processed when appraising the bids before a decision is made. The only additional costs involved arise from doing this in a consistent, well-documented manner and in somehow making this data available. Recording all of this detail would involve a small amount of expenditure but does take time at a point when officials are usually under pressure to make decisions and start spending money.

Unfortunately, it is arguable that costs are not the major barrier in terms of data availability. Using policy design to assess causal effects ideally requires government to have detailed information about the decision-making process. How were bids solicited and assessed? How were the winning bids selected? How were funding levels decided? At least in the context with which I am most familiar (the UK), it is remarkable how little of this information is systematically recorded even for internal purposes. I would assume that this problem applies much more widely beyond the UK. Assuming all this information (on the policy process and outcomes) is available, there is one remaining major barrier. Specifically, effective policy evaluation needs the government to make all this information available to researchers. For all kinds of reasons, governments remain reluctant to do this.

Of course, a genuine reason for resistance to transparency is that some of the information may be confidential (more so when it relates to individuals or firms

than areas). Fortunately, government departments and statistical agencies do appear to be increasingly willing to find mechanisms for circumventing this specific problem. In the UK, for example, they do this by making data “publicly” available to use in a secure data environment with controlled access and detailed disclosure rules (e.g., the ESRC-funded Secure Data Service). Again, there will be some cost to maintaining this data and providing access to it. The final barrier to more careful policy evaluation is that government needs to be patient. To perform the kind of analysis discussed in this section requires data on the policy and for a range of outcome variables, for example, firm performance, employment, and unemployment, for an appropriate number of geographical areas. That outcome data is usually only available with a time lag of several years which complicates the interaction between evaluation and policy formulation (because policy makers are often working on shorter time scales). But once the data becomes available, if the policy design is such as to interest academics, researchers will then spend many (unpaid) hours figuring out whether the policy in question had any causal impact on outcomes. In short, with a little patience and transparency, open evaluation has the scope to significantly increase our understanding of the causal impact of government urban policy at very little (direct) financial cost. In addition, such evaluation can also increase our understanding of how the spatial economy functions. For example, evaluation of place-specific policies can tell us the extent to which other “amenities” are likely to get capitalized into land values. Policy evaluation of transport projects can tell us whether or not market access (through the transport network) affects productivity. Looking at the impact of training policies can help increase our understanding of local labor markets. The geographical economic literature is only just beginning to explore these issues, but experience from other fields suggest that we might learn a lot more from such an approach.

35.5 Conclusions

To some extent, this chapter has been concerned with the “process” by which geographical economics influences policies. It has considered a number of barriers that limit this influence. The chapter has been structured around the three sets of constraints facing academic researchers – specifically in terms of the availability of data, the limitations of spatial analysis, and the role of the evaluation of government policy. But along the way, the chapter has also highlighted a number of constraints facing policy makers. Policy makers are accountable for the performance of particular places. This means that they need to be interested in data for administrative units even if they understand that these might not adequately capture how spatial disparities are developing and what, if any, impact policy is having. A lack of analytical capacity often exacerbates the problems caused by any disconnect between the data used for analysis and that used to assess the performance of different administrative units. In terms of the analysis, policy makers often perceive ethical or political problems with decision-making processes, such as randomization or competitive bidding, which many researchers advocate as “ideal” for

evaluation purposes. More careful evaluation calls for up-front costs in terms of systematic data collection but only delivers longer-term results once outcome data becomes available and analysis has been undertaken. Political imperatives, for example, an incentive to show short-term results, can easily override the desire of officials to take a longer-term view of the impact of the policies for which they are responsible.

Some of these issues stem from fundamental conflicts of interest between researchers and policy makers. Others are more easily addressed. Collecting data for more “sensible” spatial units – such as metropolitan areas – can better align the spatial scales used by policy makers and analysts. Using institutional features of policies to help improve the understanding of the causes of spatial disparities increases the relevance of academic research to the policy-making community. Secure data services allow governments to share data in a way that maintains some control over exactly how that data is used. In turn, open data allows for open evaluation where the academic community can provide longer-term assessments of the impact of policy even if policy makers’ attention remains focused on the short term.

Of course, addressing all of these barriers is only a necessary, but not sufficient, step in ensuring that insights from geographical economics help inform spatial policy. Belief- or principle-based policy making still trumps evidenced-based policy making in many situations. But addressing these problems also makes for good geographical economics regardless of any influence on policy. Fortunately, for academic researchers, even if we fail to change the world, improving our understanding of how the world works is hopefully reward enough for our efforts.

References

- Angrist J, Pischke JS (2009) *Mostly harmless econometrics*. Princeton University Press, Princeton
- Baldwin R, Forslid R, Martin P, Ottaviano GM, Robert-Nicoud F (2003) *Economic geography and public policy*. Princeton, Princeton University Press
- Banerjee A, Duflo E (2009) The experimental approach to development economics. *Annu Rev Econ* 1:151–178
- Baum-Snow N, Marion J (2009) The effects of low income housing tax credit developments on neighbourhoods. *J Public Econ* 93:654–666
- Busso M, Gregory J, Kline P (2010) Assessing the incidence and efficiency of a prominent place-based policy. NBER Working Paper #16096
- Cheshire P, Gordon I, Gibbons S (2008) Policies for mixed communities: a critical evaluation. Spatial Economics Research Centre Policy Paper #002
- Cheshire P, Magrini S (2009) Urban growth drivers in a Europe of sticky people and implicit boundaries. *J Econ Geogr* 9:85–115
- Combes PP, Duranton G, Overman HG (2005) Agglomeration and the adjustment of the spatial economy. *Pap Reg Sci* 84:311–349
- Dachis B, Duranton G, Turner M (2011) The effects of land transfer taxes on real estate markets: evidence from a natural experiment in Toronto. *J Econ Geogr* 12:327–354
- Department of Communities and Local Government (2006) State of the English Cities. <http://webarchive.nationalarchives.gov.uk/20070108123845/http://odpm.gov.uk/index.asp?id=1163940>

- DiNardo J, Lee DS (2010) Program evaluation and research designs. In: Ashenfelter O, Card D (eds) *Handbook of labor economics*, vol 4. Elsevier, Amsterdam
- Duranton G (2011) California dreamin': the feeble case for cluster policies. *Rev Econ Anal* 3:3–45
- Duranton G, Gobillon L, Overman HG (2011) Assessing the effects of local taxation using microgeographic data. *Econ J* 121:1017–1046
- Duranton G, Overman HG (2005) Testing for localisation using micro geographic data. *Rev Econ Stud* 72:1077–1106
- Gibbons S, Overman HG (2012) Mostly pointless spatial econometrics. *J Reg Sci* 52:172–191
- Krugman P (1991) Increasing returns and economic geography. *J Polit Econ* 99:483–499
- Lee DS, Lemieux T (2010) Regression discontinuity designs in economics. *J Econ Lit* 48:281–355
- Manski CF (1993) Identification of endogenous social effects: the reflection problem. *Rev Econ Stud* 60:531–542
- Markusen A (2003) Fuzzy concepts, scanty evidence, policy distance: the case for rigour and policy relevance in critical regional studies. *Reg Stud* 37:701–717
- Martin R, Sunley PJ (2011) The new economic geography and policy relevance. *J Econ Geogr* 11:357–370
- Neumark D, Kolko J (2010) Do enterprise zones create jobs? Evidence from California's enterprise zone program. *J Urban Econ* 68:1–19
- Overman H (2010) "GIS a job": what use geographical information systems in spatial economics. *J Reg Sci* 50:165–180
- Sacerdote B (2001) Peer effects with random assignment: results for Dartmouth roommates. *Q J Econ* 116:681–704

Section V

Location and Interaction

market
economic

sellers
traffic
modeling

choice
interregional

complexity
activities

time
locations

travel

equilibrium
social

supply
car

migration

region

land-use

transportation
onlocation

input
public

output

capital

labor

trips

industry solution

products

gold

network

black

white

gray

red

blue

green

purple

orange

yellow

pink

destination
behavior

black

white

gray

red

blue

green

purple

orange

yellow

pink

cost
route
space
accessibility

black

white

gray

red

blue

green

purple

orange

yellow

pink

Kenneth Button

Contents

36.1	Introduction	686
36.2	The Behavior of Individuals	686
36.3	Modeling Travel Behavior and Demand	687
36.4	The Elasticity of Travel Demand	693
36.5	Using Travel Behavior and Travel Demand Information	697
36.6	Conclusions	701
	References	702

Abstract

This chapter focuses on the ways in which travel behavior and demand are analyzed within the framework of regional science. Unlike numerous recent surveys that cover the more technical and abstract aspects of mathematically modeling travel behavior and demand, the attention here is more on the practical aspect of applying travel behavior and demand analysis to subjects such as regional development, infrastructure investment, and congestion analysis. Thus, while the main methods of modeling travel behavior and demand are outlined and critiqued, there is also considerable references to such things as demand elasticities and their estimation that are at the core of applied regional analysis. These types of parameter provide a direct link between a soft policy shift or a harder infrastructure investment, travel behavior, and ultimately the implications of this for regions. There is also discussion of the uses made of the forecasts that are the de facto rationale for studying travel behavior and travel demand, and the ways that neutral forecasting can be manipulated in decision-making.

K. Button

School of Public Policy, George Mason University, MS-3B1, Arlington, VA, USA
e-mail: kbutton@gmu.edu

36.1 Introduction

Whatever we do takes up time; it takes time to watch a movie, to appreciate a good meal, or to see a favored soccer team lose. Time is important; why else do fast-food restaurants flourish? We like to save time to allow us to do more things over our limited life spans and in particular because there is considerable uncertainty about its duration. In a more prosaic context, we often try to save time to increase our economic productivity, or perhaps our employers do, because, after all as the old proverb says, “time is money.”

Time-saving is important in influencing travel behavior and on travel demand, as in most other activities, sometimes more so and sometimes less. The relationship is, as in all other similar matters, a complicated one because from a behavioral perspective, time is never consumed in isolation. Just as going to the opera involves consuming time, it also means spending money. Traveling thus costs money in several ways, be it a fare, fuel, or shoe leather, and involves some form of final consumption, and not just at the end of a journey, although this is often emphasized. In this sense, the demand for travel is also derived from what is to be “consumed” at the end of a trip, be it a work or leisure final activity; in economic terms time is a joint product.

Travel, therefore, entails considering the consumption not just of time but also of money and other things as well as enjoying a number of benefits at the end of the trip, in some cases, reading, sightseeing, listening to the radio, or enjoying having your Lamborghini admired on the trip. It is not surprising therefore that in practice, transportation analysts are often very bad at predicting travel behavior, particularly when there are major changes in underlying parameters over time or when some of the costs and benefits are not easily quantified.

36.2 The Behavior of Individuals

The focus here is exclusively on the travel behaviors of individuals; it does not consider the movement of goods or information electronically, other than those in cases when these may impact on personal movement. The overlap can, however, be rather more extensive than is often reflected in policy-making and academic analysis. In most cases, individuals accompany goods when they move, trucks would be immobile without truck drivers, and the personal involved have their own individual traits. In many cases, goods movements use the same infrastructure as individuals or the same piece of mobile plant, for example, passengers on the top deck of an aircraft and freight in the belly hold. The electronic transportation of information is important when it substitutes for a personal movement, for example, in the context of teleworking, or when it facilitates personal travel, as with airline computer booking systems or with automobile route guidance systems. We also say relatively little about where origins and destinations of trips are located, but there are clear links between where people choose to live and work and transportation facilities that in turn feed back on their travel behavior.

Additionally, handbooks have many purposes and many potential readerships ranging from those concerned with the highly theoretical and abstract to those wanting a quick guide to finding a practical, rough-and-ready solution to an immediate real-world problem. (Volumes dealing with the latter are often, with a degree of derision, described as manuals, but their approach is no different to “handbooks” providing recipes for developing abstract theorems.) Here we seek to function in the middle ground, by both discussing many of the widely used practical approaches to analyzing travel behavior and travel demand and making use of our knowledge of the subject, but also to offer some guidance as to the direction research at a more theoretical level is moving. While there will be some discussion of model derivation, in the spirit of handbooks, the emphasis is on setting down what we have and is currently being used, rather than plowing through all of the intellectual and mechanical background in detail.

36.3 Modeling Travel Behavior and Demand

Travel behavior is like any other activity, possibly excluding religion and politics, in that rational economic forces largely drive it. People base their decisions on the benefits they will enjoy from it, either directly or at the end of the trip, constrained by the generalized costs of the movement in terms of time and money, relative to available resources and the opportunity costs of using them in some other activity. Put this way, the study of travel behavior may seem rather trite; it is basically a constrained maximization problem. The devil, as is often the case, lies in the detail.

At the outset, it is important to distinguish between travel behavior and travel demand. The latter is a particular influence on the former. Travel behavior is what people actually do, the way that they behave, the trips they make, and the forms of transportation that they use. It is basically what the collection of transportation data on person and goods movements reflects and what those who invest in transportation or manage transportation assets try to forecast. Travel demand is one part of this and reflects what people’s travel behavior would be with various forms of transportation facilities available. It makes no allowance for the roads or transit facilities that exist, other than to extrapolate in some cases their current use to forecast future travel behavior. Thus, while travel behavior is dependent on travel demand, it is not only the demand for travel that influences final travel behavior outcomes but also the supply of transportation facilities.

Travel behavior can also be very variable in its nature, and virtually all analysis has to be taken as contextual and in particular is related to the physical facilities available and the flexibility the traveler enjoys. The timing of trips is far more important for commutes, for example, than for daily leisure activities, although over a longer period, the constraints of having set periods for vacations can be very restrictive in terms of vacation travel. Equally, travel at peak times of the day, the “rush hour,” or at common vacation times, such as Christmas, can put pressures on the existing transportation infrastructure, with those able trying to avoid the worst

excesses of congestion. But the very transport demands of people also affect the de facto supply of capacity available for their individual use. The classic example is that economies of agglomeration that, by concentrating employment, tend to push up the demand for urban transport infrastructure use in the morning and evening rush hours, leading to congestion and the increasing of the prices of these trips. This “club-good” problem, akin to golfers preferring to tee off in the early morning, puts pressure on the transport system leading to high levels of peak period congestion, a phenomenon that has attracted a lot of attention among urban and transportation scientists (Lindsey and Verhoef, 2008).

The information traditionally used for analyzing travel behavior and demand has been of the revealed preference form, and the tools of analysis have been Gaussian in orientation. Essentially this has involved the extrapolation of previous travel demand behavior into the future based on statistical analysis of prior relationships between travel and physical, economic, and sociological influences. The relationships are assumed to be constant over time. The stochastic nature of the underlying historical relationship allows confidence intervals to be drawn around the projections.

The traditional modeling approach to handling this information to, for example, forecast the implications on travel behavior of a policy shift in road investment was originally limited in part by the need for computational convenience. The models were developed mainly in the 1960s at a time when large-scale transportation and land-use plans were in vogue with an emphasis on providing road access along major commuter corridors, and this influenced the types of forecasts being sought. But these plans required considerable details of traffic effects over large interconnected networks, which in turn necessitate complex manipulating of very large databases.

The methodology involved breaking down travel demand analysis into a number of sub-questions, the “four-stage model.” In the urban context, this involves, what is essentially recursive modeling of the aggregate travel in an area, disaggregating this into trips between areas with the city, disaggregating these trip distributions according to the modes used, and finally assigning traffic to individual routes. In simple mathematical terms, the stages can be expressed as

$$T_i = f(X_i); \quad T_j = f(X_j) \quad (36.1)$$

$$T_{ij} = f(T_i, T_j, C_{ij}) \quad (36.2)$$

$$T_{ijm} = f(T_{ij}, C_{ijm}, C_{ijm'}) \quad (36.3)$$

$$T_{ijmp} = f(T_{ijmp}, C_{ijmp}, C_{ijmp'}) \quad (36.4)$$

where T_i is the number of trips originating in i , T_j is the number of trips destined for j , T_{ij} is trips between i and j , X_i and X_j are the socioeconomic features of i and j , T_{ijm} is the trips between i and j by mode m , T_{ijmp} is the trips between i and j by mode m

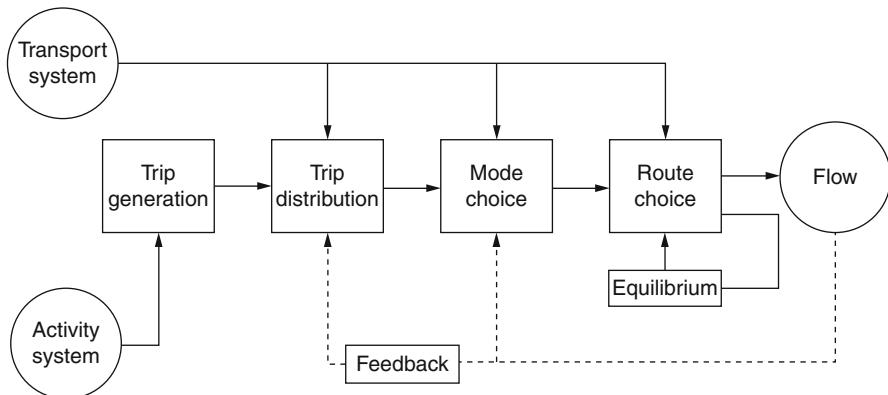


Fig. 36.1 The four-stage model sequence

along route p , C_{ij} is the generalized cost of travel between zones i and j , C_{ijm} is the generalized cost of travel between zones i and j by mode m , C_{ijmp} is the generalized cost of travel between zones i and j by mode m using route p , and the prime notation refers to alternative modes (m') or routes (p'). The four stages and their links with inputs, standard feedbacks, and outcomes are seen in Fig. 36.1.

This highly aggregated approach that looks at travel behavior in terms of zonal flows, however, has limited behavioral content. It also suffers from a number of technical weaknesses both in terms of the overall model and of individual submodels, for example, it is recursive by nature, and although feedback loops are possible to reflect the impacts of traffic assignment on aggregate travel, this tends to be a mechanical rather than a behavioral process. It is also difficult to assess the overall statistical fit of the model; a series of relative small errors in individual components could be compounded over the sequence.

Despite these limitations, these four-stage models are still widely used for land-use transportation planning exercises, in part because they are relatively easy to understand and software is abundant, but intellectually they have been superseded by approaches more embedded in economic and social sciences and in particular by discrete choice and activity-based models.

Broadly, disaggregate models are characterized by two main features. First, they explicitly recognize that travel decisions emerge out of individuals' optimizing behavior and, if the final goods consumed as a result of travel are normal, then at a very minimum the demand for travel ought to be related positively to disposable incomes and negatively to the prices of transportation services. Secondly, most have their origins in the "attribute theory of demand." This approach to human behavior assumes that people desire to maximize a utility function that has, as its arguments, commodity attributes rather than the quantities of the actual goods consumed. In other words, if we represent the amounts of attributes by the vector z , the amounts of commodities (in this case travel alternatives) by the vector x , posit a utility function, $U(z)$, and a production of attribute function, $G(x)$, which reflects

the attributes of different travel alternatives, and assume that potential travelers are constrained by income, y , and the price of travel, p , then we can reduce the problem to solving

$$\begin{aligned} & \max U(z) \\ \text{subject to } & z = G(x) \\ & x \geq 0 \\ & px \leq y \end{aligned} \tag{36.5}$$

Because the unit of analysis is the household or individual, the decision to make a particular trip or use a specific mode requires some form of a discrete model specifications. Depending on the nature of the case, these are normally derivatives of logit or probit models; for estimation purposes, this can generally be expressed in log-odd terms as $\ln\left(\frac{P_i}{1-P_i}\right) = v(x_i)$, where P_i is the probability of, say, trip i being made by automobile given the attributes, x , of car travel and alternative modes. The framework can be extended, for example, to situations when there are multiple concurrent choices (multinomial logit model) or when choices are sequential (nested logit models). Theoretically, these types of model have a firm base in economic science and on the idea of random utility developed by McFadden (1974). The models have the practical advantage of requiring relatively small data sets for estimation, and there is generic econometric software available to handle discrete choice situations. What they do not do is provide a mechanism for looking at large-scale shifts in travel demand across a large area or for integrating travel behavior with wider changes in activity patterns.

While the sequential and disaggregate approaches to transportation demand analysis concentrate on developing sophisticated mathematical simulations of travel behavior, recently there has been a growth of interest in “behavioral realism” and an emphasis on “understanding the phenomenon.” Sometimes called the activity-based approach because it has sought to embrace a richer, holistic framework in which travel behavior is analyzed as a daily or multiday pattern of behavior related to lifestyles and participation in various activities.

The idea is that the demand for travel is derived from the activity patterns of individuals and households. The basic idea has much to do with the concept of time geography dating back to Hägerstrand (1969) in which an individual’s choice of a specific activity pattern is viewed as being the solution to an allocation problem involving limited resources of time and space. In this sense, simply focusing on actual behavior is not that useful, but rather there is a need to put more emphasis on the constraints that limit people’s behavior, and these are often more difficult to define and measure.

These approaches to modeling have also been tied in with the greater use of stated preference techniques that question people about their probable travel reactions to say a change in gasoline prices or the introduction of a new public transportation service, rather than consider the revealed preferences of people to similar changes in the past. Stated preference methods themselves can be applied to

most forms of behavioral analysis, including narrower trip-based work, and has been more strictly defined by Kroes and Sheldon's (1988) as "a family of techniques which use individual respondents' statements about their preferences in a set of travel options to estimate utility functions," and is often claimed to be helpful when:

- There is insufficient variation in revealed preference data to examine all variables of interest.
- There is a high level of correlation between the explanatory variables in a revealed preference model making statistical estimation of parameters difficult.
- A radically new technology or policy takes the analysis outside of the realms where current revealed behavior is relevant.
- Variables are not easily expressed in standard units (e.g., when the interest is in the effects on demand of less-turbulent travel by air).

The aim of activity-based modeling, a specific form of interactive modeling, is to develop models that get closer to the essential decision process underlying travel behavior, and it is for this reason that stated preference techniques have generally been favored in this type of work. Rather than simply incorporate variables such as household status in mathematical models because the statistical "explanation" of the model appears to be improved, activity modeling seeks to explain why status affects travel behavior. Theoretically, travel is seen as one of a whole range of complementary and competitive activities operating in a sequence of events in time and space. It is seen to represent the method by which people trade time to move location in order to enjoy successive activities. Generally, time and space constraints are thought to limit the choices of activities open to individuals. The technique is still far from fully developed, but it began to be applied in a relatively limited number of small-scale forecasting studies from the late 1970s.

The emphasis of activity-based models is upon the household (or individual) as the decision-making unit. It focuses on the revealed pattern of behavior represented by travel and activities, both in home and nonhome over a defined period of time. It thus generally makes use of revealed preferences but can combine this with stated preferences analyses for forecasting the impacts of changes in activity options.

According to Heggie (1978), ideally, an activity-based model should exhibit six main properties:

- It should involve the entire household and allow for interaction between its members.
- It should make existing constraints on household behavior quite explicit.
- It should start from the household's existing pattern of behavior.
- It should work by confronting the household with realistic changes in its travel environment and allowing it to respond realistically.
- It should allow for the influence of long-term adaptation.
- It should be able to tell the investigator something fundamental that he did not know before.

In general, the approach is typified by a fairly small sample and careful survey techniques, often involving such things as "board games" – such as the

“household activities travel simulator” (HATS) developed by the Oxford University Transportation Studies Unit (Jones, 1978) – or other visual aids, frequently computer based these days, to permit households to appreciate the full implications of changes in transportation policy for their own behavior. In a sense, it represents an attempt to conduct laboratory experiments by eliciting responses in the context of known information and constraints.

The early HATS approach was to confront a household with a map of the local area together with a 24-h “strip representation of colored pieces” showing how current activities of the household are spread over space and throughout the day. Changes to the transportation system were then postulated and the effects on the household’s activities throughout the day were simulated by adjustments to the strip representation. In this way, changes in the transportation system could be seen to influence the 24-h life pattern of the household, and apparently unsuspected changes in “remote” trip-making behavior can be traced back to the primary change. It makes clear the constraints and linkages that may affect activity and transportation choices. More recent studies have adopted rather more sophisticated experimentation procedures, often involving computers, which provide for greater flexibility and easier interaction with those being “interviewed” – for a survey, see McNally and Rindt (2008). Examples of programs of this genre include ALBATROS (A Learning-Based Transportation Oriented Simulation System), the first computational process model of the complete activity scheduling process that could be fully estimated from data, and TRANSIMS, an attempt to replace the entire traditional travel paradigm, has an activity-based front end linked to a population synthesizer, and integrated with a micro-simulation of modeled travel behavior. Computer-based models include STARCHILD and AMOS with mathematical programming models such as HAPP. The development of geographical information and global positional systems has allowed for better data availability and real-time surveillance of travel and associated behavior.

While this aspect of the approach has been refined, important technical issues remain regarding using the information gathered from stated preference type experiments for forecasting. There is still, for example, much to be learned about why some households give strategically biased responses; in particular there are difficulties in handling habit, inertia, and hysteresis in an experimental framework. At a more technical level, John Bates (1988) points to our lack of knowledge about the error structures associated with stated preference data and the particular problems of pooling data across individuals.

In contrast to the more traditional revealed preference schools, advocates of this approach, however, point to both the specific recognition that travel is a derived demand and the fact that transportation policies have qualitative, as well as quantitative, effects on people’s lives. In the longer term, when operational models are more fully developed, the framework may offer the much-sought-after basis for integrating land-use and transportation planning assessment. In the short term, the approach has offered useful insights and a method for cross-checking the validity of conventional statistical analysis of behavioral data.

36.4 The Elasticity of Travel Demand

In many situations, the focal point for travel demand analysis is very specific, relating to a particular issue or policy question such as the impact of a fare rise on transit ridership or the implications for travelers of a new emissions charge. While larger models are often used to assess the general effects of these types of change, the estimation of the appropriate demand elasticity provides guidance to the specific effects on a target variable. There are an entire range of possible elasticities that may be considered (Oum et al., 2008), but all have the common feature of measuring the proportional change in a behavior relating to a target transportation variable (bus trips, fuel use, car pooling, or whatever) of a proportional change in the instrumental variable (a new tax, a new vehicle design regulation, the reduction of a speed limit, or whatever).

The estimation of basic elasticities is conceptually fairly straightforward (at its simplest it is $\varepsilon_A = \{\delta Q/Q\}/\{\delta P/P\}$ where Q is the “quantity” of travel measured in some way and P is the “price”) and spelled out in introductory microeconomics texts. For example, if we take a travel demand function, say, for bus travel, of the log-linear form $\ln Q_M = \alpha + \beta_1 \ln P_M + \beta_2 \ln Y + \beta_3 \ln P_N$, where, Q_M is the quantity of bus services demanded, P_M is the fare for bus travel M , Y is the income, and P_N is the price of an alternative, say, car travel, then N , the fare elasticity is parameter β_1 with the income elasticity of demand, reflecting the sensitivity of the quantity demanded to income changes being β_2 and the cross-elasticity of demand regarding the cost of car travel being β_3 . Here we focus on some of the more transportation specific nuances and also offer some discussion of the empirical values that have been derived.

Generalizations about the size of elasticities are difficult, especially across all modes of transportation, but in many cases it seems clear that price changes within certain limits have relatively little effect on the quantity of travel or transportation services demanded. Further, while demand elasticities often exhibit a degree of stability over time, they do not remain constant in part because of shifts in the demand function due to such things as rising incomes or changes in consumer tastes. Studies of urban public transportation in the 1970s, for example, covering a variety of countries indicate relatively low price elasticities with a direct fare elasticity of around -0.3 being considered normal, but emerged as somewhat higher in the 1980s. The effect of price change on private car travel must be divided between the effect on vehicle ownership and that specifically on vehicle use. Most early United Kingdom studies of car ownership, for example, indicated an elasticity of about -0.3 with respect to vehicle price and -0.1 with respect to gasoline price, but empirical work suggests a rather higher sensitivity in the United States: -0.88 purchase price and -0.82 fuel price elasticities. The generally low fuel price elasticity for car use in the short term is attributable to changing patterns of household expenditure between vehicle ownership and use and people’s perception of motoring costs. Bendtsen (1980) brought early findings together in an international comparison that found the petrol price elasticity of demand for car use to be -0.08 in Australia for the period from 1955 to 1976, -0.07 in Britain for

Table 36.1 Price elasticities of demand for passenger transportation expressed in absolute values

Mode	Range surveyed		Number of studies
	Market demand elasticities	Mode choice elasticities	
Air			
Vacation	0.40–4.60	0.38	8
Non-vacation	0.08–4.18	0.18	6
Mixed ⁺	0.44–4.51	0.26–5.26	14
Rail: intercity			
Leisure	1.40	1.20	2
Business	0.70	0.57	2
Mixed ⁺	0.11–1.54	0.86–1.14	8
Rail: intracity			
Peak	0.15	0.22–0.25	2
Off peak	1.00	n.a.	1
All day ⁺	0.12–1.80	0.08–0.75	4
Automobile:			
Peak	0.12–0.49	0.02–2.69	9
Off peak	0.06–0.88	0.16–0.96	6
All day ⁺	0.00–0.52	0.01–1.26	7
Bus:			
Peak	0.05–0.40 ⁺⁺	0.04–0.58	7
Off peak	1.08–1.54	0.01–0.69	3
All day ⁺	0.10–1.62	0.03–0.70	11
Transit System:			
Peak ⁺⁺⁺	0.00–0.32	0.1	5
Off peak	0.32–1.00	n.a.	3
All day ⁺	0.01–0.96	n.a.	15

Source: Oum et al. (2008)

1973/1974, –0.08 in Denmark for 1973/1974 and –0.12 for 1979/1980, and –0.05 in the United States for the period from 1968 to 1975. Oum et al. (1992) found a slightly greater degree of sensitivity when looking at seven studies covering the United Kingdom, the United States, and Australia; they yielded car usage elasticities in the range –0.09 to –0.52.

Table 36.1 provides a survey of some estimates of automobile and public transportation elasticities. The market demand elasticities reflect the impacts on total mileage of a price change, while the mode choice elasticities refer to the probabilities of using a mode as its price varies. We immediately note that the former combines mode shift and distances per trip and thus tend to be greater than mode choice elasticities. Further, we see a wide range of elasticities emerging dependent on the nature of the trip, the mode used, and the time of day the travel takes place. In particular, when the timing and need for trips is very rigid, such as journeys to work during peak periods, the elasticities tend to be quite low. Leisure travel, where there is often more flexibility, is more sensitive to travel costs.

There is, in addition to the data in the table, an abundance of evidence that the fare elasticity for certain types of public transport trips is much higher than others. Business travel demand in particular seems to be relatively more insensitive to changes in transportation price than other forms of trip. The pioneering work of Kraft and Domenich (1970) found that public transportation work trips exhibited a fare elasticity of -0.17 in Boston compared with -0.32 for shopping trips. A similar pattern is found for business and nonbusiness air travel with the latter being generally higher. This is a pattern also seen, as one might expect, in terms of the type of fare being paid. Straszheim (1978) found some time ago that, "First class fares can be raised and will increase revenue... The {price elasticity of} demand for standard economy service is about unity, and highest for peak period travel ... The demand for discount and promotional fares is highly price elastic...." This conforms to the intuition that vacation travelers have more flexibility in their actions (destinations, times of flights), whereas business trips often have to be taken at short notice. The lower sensitivity associated with first-/business class fares also reflects the service requirements of users who often seek room to work on planes and lounges. The estimates of the elasticities are also sensitive to the length of service with shorter routes generally exhibiting higher fare elasticities, in part because other modes of transportation become viable options.

Users of different forms of transportation, or different services of the same mode, are often confronted with a variety of payment options; their perceptions of the price of a journey may differ from the actual monies expended. Motorists generally perceive very little of the true overall price of these trips because they base decisions on a limited concept of short-term marginal cost – for example, they only buy fuel periodically and do not take this into account when deciding whether to make a particular trip – whereas users of public transportation have to buy tickets before traveling making them more aware of the costs of their behavior. Nevertheless, because of the range of season tickets that permit bulk buying of journeys over a specific route, and travel card facilities that permit bulk buying of journeys over a specified network, the distinction is not a firm one.

As with other purchasing decisions, people confronted with a change in transportation price generally act differently in the ultrashort run, the "market period," the short run, and the long run – [Table 36.1](#) offered some examples. Immediate reactions, in the ultrashort term, to a public transportation fare rise may, for example, be dramatic, with people, almost on principle, making far less use of services, even boycotting it, but knee-jerk reaction is extremely short lived and seldom considered by economists, although it is often of interest to politicians. This behavior is usually short lived. In the slightly longer "market period," people revert to their initial behavior relatively unresponsive to a price change either because they do not consider it a permanent change or because technical constraints limit their immediate actions; the elasticity is virtually zero. Over time, people can adjust their behavior and, in the short run can change their travel patterns by switching modes, combining trips, and cutting out some travel, and businesses can reschedule the use of their vehicle fleets and modify their collection and delivery patterns. The demand for cars travel, therefore, becomes more elastic in

relation to the new price. In the long term, people can change the type of car they use and their employment and residential locations, and industry can modify their entire supply chain. Taking a specific context, when considering the effect of general rises on commuter travel costs, the necessity of having to make journeys to work is likely to result in minimal changes in travel patterns in the short term, but over a longer period, relocations of either residence or employment may produce a more dramatic effect. This implies that one must take care when assessing elasticity coefficients and it is useful to remember that cross-sectional studies tend to offer estimates of long-run elasticity while time-series studies reflect short-term responses.

Elasticities are also generally found to increase the longer the journey under consideration. This is not simply a function of distance but rather a reflection of the absolute magnitude of, say, a 10 % rise on a \$5 fare compared with that on a \$500 fare. It is also true that longer journeys are made less frequently, and thus, people gather information about prices in a different way. Additionally, they often tend to involve leisure rather than business travel; this suggests that distance may be picking up variations in trip purpose. In the air transportation market, for example, DeVany (1974) found in a classic study that price elasticity rose from -0.97 for a 440-mile trip in the United States to -1.13 for an 830-mile trip.

Turning to the effects of income on travel behavior, while there is ample evidence that travel is a normal good in the sense that more is demanded at higher levels of income, this generalization does not apply to all modes of travel or to all situations. At the national level, income exerts a positive influence over car use, but this is not so clear-cut with public transportation use, and in some cases the latter becomes an “inferior good” with its use falling after some level of income has been attained. Gwilliam and Mackie (1975), for example, suggest that the long-run elasticity of demand with respect to income was of the order -0.4 to -1.0 for urban public transportation trip making in the United Kingdom. They argue that although car ownership rises with income, and hence some trips are diverted from public transportation, there is still a limited offsetting effect inasmuch as wealthier households make more trips in total. This effect would seem to be less relevant today with much higher levels of automobile ownership in developed countries, a fact borne out in the findings of Crôte et al. (2009) in the context of Mexico.

The income elasticity of demand for many other modes of transportation is seen to be relatively high, and especially so for modes such as air transport. Taplin (1980), for example, suggests a figure of the order of 2.1 for vacation air trips overseas from Australia. By its nature, air travel is a high-cost activity with the absolute costs involved being high even where mileage rates are low so that income elasticities of this level are to be expected. There is also some evidence that wealth influences the demand for air travel, with Alperovich and Machnes's (1994) study of the Israeli market finding a wealth elasticity of 2.06.

As with price, income changes exert somewhat different pressures on travel behavior in the long run compared with the short. In general, a fall in income produces a relatively dramatic fall in the level of demand, but as people readjust their expenditure patterns in the long term, the elasticity is likely to be much lower.

Reza and Spiro (1979), for example, produce an estimate of 0.6 for the short-run income elasticity of demand for petrol rising to 1.44 in the long run. If one assumes that gasoline consumption is a proxy for trip making, then one could attempt to justify this in terms of a slow reaction to changing financial circumstances a reluctance, for example, to accept immediately the consequences of a fall in income. In fact, the situation is likely to be more complex because the long run may embrace changes in technology, and possibly locations, that alter the fuel consumption-trip-making relationship. Thus, these figures may still be consistent with the initial hypothesis regarding the relative size of short- and long-run income elasticities of demand for travel.

The demand for any particular travel service is likely to be influenced by the actions of competitive and complementary suppliers – the cross-elasticity effect. Strictly speaking, it is also influenced by prices in all other markets in the economy that touch upon the importance of motoring costs vis-à-vis the demand for public transportation services. There are wide variations in results that generally reflect the adoption of alternative estimation procedures and time-lag allowances, as well as the peculiarities of the local travel situation. One of the more interesting points is the almost total insensitivity of the demand for urban car use to the fare levels of both bus and rail public transportation modes. This fact, which has been observed in virtually all studies of urban public transportation, is the main reason that attempts by city transportation authorities to reduce or contain car travel by subsidizing public transportation fares have, in the main, proved unsuccessful.

Transportation demand is also sensitive to the quality of service offered, although measurement of relevant elasticities is challenging because many service attributes are essentially intangible and because of this are treated more as service qualitative. It is noticeable, for example, from empirical studies that public transportation demand is sensitive to changes in service quality attributes such as reductions in speed or frequency of services; other attributes that are often intuitively seen as important are not easily isolated. Lago et al. (1981) examined a wide range of international studies concerned with urban public transportation service elasticities and concluded that increased service levels do not generate proportional increases in passengers and revenues, but their analysis looked at service quality attributes in isolation rather than at a package of service features and missed many qualitative attributes. The survey also highlights that service quality is far more important when the initial level of service is poor; the general elasticities found for peak period ridership, for instance, are much lower than those for the off peak. Further, it found that the service headway is one of the more important service variables; the studies examined indicate an elasticity of -0.42 compared with -0.29 for in-vehicle bus travel time.

36.5 Using Travel Behavior and Travel Demand Information

Now we deviate a little from what is often contained in previous handbooks that are concerned with the “science” of analyzing travel behavior, and spend some time on

more practical matters of how research on travel behavior and demand is often applied. This is very much in the realm of regional political science, but germane to the context in which more positive aspects of the science of travel behavior and demand are treated. The implicit view of Milton Friedman (1953) on model performance can be recalled here, “The ultimate goal of a positive science is the development of a ‘theory’ or ‘hypothesis’ that yields valid and meaningful (i.e., not truistic) predictions about phenomena not yet observed.” Simply producing a model of travel behavior that tells stories about the past is unlikely to be either useful or good science.

While there are numerous studies and texts on academic modeling of travel behavior in reality, the importance of this for regional science is just how useful the analysis is for forecasting. While there may be a historic interest in knowing why particular land-use patterns have emerged or why particular local industries have thrived, the main rationale for studying travel behavior and travel demand is to be able to use the information gleaned for policy development. This means understanding not only why the current demand patterns exist but also how they may change in the future. Most existing models can explain ongoing trends fairly well, and the importance of trend breaks to some extent but are poor at explaining new trends. The advent of real-time information systems and mobile communication platforms has, for example, added to the way people view trip making and the speed at which they change their travel plans. The switch to large-scale service sector employment is changing the perceptions of working hours. These were developments not foreseen 20 or so years ago when much of the thinking surrounding the then current travel behavior and demand models was emerging.

One of the main purposes of trying to get a handle on travel behavior and demand is to assist in policy-making, both in the public and the private sector. While there is a plethora of academic interest in trying to improve our understanding of why people travel and the nature of their trips when they do go into motion, there is somewhat less study of travel behavior by the private sector that provides hardware such as rail track and automobiles and software such as insurance or those that cater for some of the side effects of travel, such as medical services. Just reflecting on the last of these, changes in modes of road travel between public and private means affect the types of injuries associated with accidents and the incidents of ailments such as severe asthma. In terms of the automobile industry, the differing driving patterns of various age and income groups affect the demands for their models, as do social attitudes toward various modes of transport.

Aside from those directly involved in transportation, there are others with an interest in travel behavior, not least of which are the fiscal authorities. Transportation is both a large generator of taxation revenue and a major sump hole for subsidies. There are also matters of the demands of transportation users such as the military that are seldom considered within conventional academic modeling or at least the material that appears in the public domain.

Each group has an interest in a particular aspect of travel behavior. Transportation and land-use planners often want a longer view to assess the implications of fixed, often multimodal, infrastructure investments and their interactions with

economic and social development. The car industry is focused just on forecasts that lead to the commercial success of its products, and this involves a somewhat shorter time horizon. Most tax authorities do not often go beyond the myopia of trying to balance this year's books.

Following the greater academic herd, we focus on the types of demand analysis used in land-use transportation planning and policy-making. Here hard numbers are generally concerning the future use of links and nodes in the relevant transportation system. These are needed in particular for engineering design purposes and in the estimation of costing to ensure appropriate funding is available.

The situation is not a very satisfactory one, and despite the efforts of analysts, the poor quality of transportation forecasts used in the field has been known for some time. The 1970s saw considerable debate in the United Kingdom, for example, over inaccuracies in the forecasting of car ownership, a major input into traffic modeling, and of the traffic forecasts themselves in the 1980s. The forecasts for the M25 London orbital road, for instance, were that, on 21 of the 26 three-lane sections, the traffic flow would be between 50,000 and 79,000 vehicles a day in the 15th year of operation, whereas the flow within a very short time was between 81,400 and 129,000.

In the 1990s, a series of studies in the United States, including those by the likes of Kain (1990) and Pickrell (1992) brought into question the forecasts of transit ridership and financial costs of investments. Pickrell's study of programs funded by the United States Urban Mass Transportation Administration, for example, found that the ten projects examined produced major underestimates of costs per passenger (e.g., the costs for the Miami heavy rail transit were 872 % of those forecast, and for Detroit's downtown people mover, they were 795 %); only the Washington heavy rail transit project experienced actual patronage more than half of that forecast. Updating of this work and looking at cost and ridership forecast for 47 United States transit systems indicates only limited improvements over time (Button et al., 2010) when allowance is made for the composition of projects (e.g., light and heavy rail) and for whether an investment was in a strictly new system or an extension of an existing one.

More recently, a series of studies by Flyvbjerg et al. (2005) looking across a range of surface transportation forecasts have shown considerable inaccuracies extending to most western economies and provide confirmation of the poor performance of forecasts. There emerges, in particular, a tendency for overprediction of capacity utilization and underprediction of the outcome costs of investments – for example, for ten rail projects from a variety of countries, the passenger forecasts overestimated traffic by 106 %, whereas for road projects, there is a tendency for the forecasts to be wrong by about 20 % but with the errors spread equally around the ultimate flows. In terms of costs, an examination of 58 rail projects indicates overruns averaging nearly 45 % and for 167 road investments, overruns of 20.4 %.

It is not just the public sector decision-making per se that is often based on poor traffic forecasts. An American study by JP Morgan (1997) of 14 privately financed and operated toll roads found that only one exceeded the projected return, with four projects over estimating forecasting returns by at least 30 %. The overall conclusion being “reducing the uncertainty associated with these forecasts represent one of the

major challenges for transportation agencies, traffic consultants, investment bankers and investors.” In a study of over 100 international, privately financed road project appraisals conducted between 2002 and 2005, Bain (2009) concludes that “...in terms of error, the predictive accuracy of traffic models – used for toll road or toll free road forecasts – is poor.” Again there emerges a proclivity to overestimate traffic flows, with the ratio of actual to forecast traffic falling below unity for the majority of studies, although in some cases the predictions underestimated by up to 51 %. The accuracy of forecasts does not appear to have improved over time, although, of course, that is not to say that some have not proved to be very reliable.

Travel behavior and demand forecasts are a major factor in both the decisions to undertake investments and in their design; serious errors in foreseeing changing patterns of travel can thus result in misuses of resources across modes and probably between travel and other activities in the economies concerned. The problem may well, in fact, be considerably worse than the quantifications by Pickrell, Flyvbjerg, and others cited earlier, in that there is no way of knowing if forecasts of investments that were rejected were biased and instrumental in the rejections. If so, then resources that would have earned a social return in those projects would have been transferred to some other use where the net benefits are less. We only have data on the investments that actually materialized and relating to policies that have actually been adopted.

The difficulties that emerge, however, are only partly a function of strict forecasting errors; there are three broad and entwined reasons why travel behavior and demand forecasts are generally poor: technical problems of the type we have discussed, carrying out the forecast, and using the forecast. The last of these is very much in line with the public choice theory of economic science that focuses on rent seeking and coalitions of interest. Thus while we have tended to couch the discussion as normative issue, it could be reexpressed in positive scientific terms along the lines of Friedman.

Each facet of the problem however is not constant, is often entwined with another, and can vary with circumstances. Regarding the first, there is a widespread proclivity to look at the technical merits of the models being used, rather than their practical use as a forecasting instrument; elegance, sophistication, and the ability to backcast are often seen as criteria for a good forecasting model, whereas in many cases, forecasting exogenous variables is more difficult than predicting traffic flows however well the model fits historical data. There are, however, clearly challenges in the collection of the data needed to calibrate the parameters of the models needed to produce forecasts and in predicting even the short-term future path of many explanatory variables. Most forecasts rely on extrapolations of previous behavior, but divergences from historical relationships do occur and are often seen in wider institutional changes reflecting social priorities and attitudes to things like the environmental impacts of transportation, its safety, and its security. The use of such techniques as sensitivity analysis and simulations can provide some insights, but the range of possibilities, and thus the potential for error, increases as the length of the forecast period gets longer.

Availability of data is always a problem for forecasting travel behavior, and the situation may be getting worse. There are certainly better survey design techniques

and more ways to gather travel-related data ranging from online surveys to the use of global positioning systems for tracking movement. There is also evidence of improvements, although some may disagree, in our ability to produce reasonable medium- and long-term forecasting of such things as income, demographics, and fuel prices that feed into travel behavior projections. Against this, the move toward lighter regulation and privatization of travel facilities has changed and often reduced the public sources of data, and especially economic data, available to forecasters to carry out their work.

Despite these largely technical and operational challenges, the problems in forecasting accuracy would often seem to lie more in the way forecasting is actually done and the way results are used. A problem initially highlighted by Kain (1990) is that forecasts and travel behavior analysis are not politically neutral, and many decisions regarding transportation investments and policy are not made in the public interest, but to some degree serve the ends of those who are making or using them. Basically, the forecasting process and output can be seen as captured by those who commission the forecasts and then make use of them. These may be politicians who wish to win reelection and thus positively assess the short-term gains of supporting high-use/low-cost investments provided by some forecasts against other less “optimistic” projections, or they may be bureaucrats concerned with increasing public sector activities. Under some forms of government, and particularly federal systems, “pork-barrel politics” can incentivize the use of biased forecasts by local agents to gain central funding: a principal-agent issue. The system is, in strict terms, corrupt and even if forecasters are neutral in their work, this has little influence on the way their output is interpreted and used. Thus, to try and close the gap between forecast and actual outcomes is often seen as a matter of institutional reform, rather than better science (Wachs, 1990).

The forecasting problem should not be seen as being unique to the public sector; there are similar institutional issues when considering some forms of private sector forecasts, especially when they involve concessions. The incentive for those tendering for such things as build-operate-transfer projects, according to Bain, is to be optimistic with forecasts so as to win the contract and gain financial support. While the evidence for this practice is largely anecdotal in the case of tolled roads, there is more support in the context of airport concessions where ex post renegotiations are relatively common when traffic flows fall short of forecasts.

36.6 Conclusions

From an economic perspective, transportation is not at all special; when deciding whether to use it, people compare its relative generalized cost against the perceived potential benefits. As with most things, people are often disappointed because the benefits of the trip do not live up to expectation or the costs include elements they had not foreseen. From the point of view of assessing the impact of policies that change travel behavior on regional economic performance or local social conditions, it is the perceptions rather than the actuality, however, that are important.

Modeling of this, despite the veneer of simple mathematics that is often used as the language of debate, is far from easy, especially in terms of providing good forecasts of travel behavior to feed into regional analysis. The situation becomes even more complicated in practice because of the ways forecasts are used is not neutral but reflects a larger “political” process that itself needs to be modeled.

The progress that has been made in terms of the pure technique of travel behavior and demand modeling has, however, been significant in terms of moving away from what was essentially seen as a mechanical, engineering process to one that embodies an acceptance that human behavior is more complex and less systematic than the early analyses assumed. It has also become more integrated into larger regional and urban modeling, with interactive relationships largely replacing the idea of a recursive structure with land-use characteristics having a one-way influence on travel behavior, activity analysis being the most important element in this.

References

- Alperovich G, Machnes Y (1994) The role of wealth in demand for international air travel. *J Transp Econ Policy* 28:163–173
- Bain R (2009) Error and optimism bias in toll road traffic forecasts. *Transportation* 36:469–482
- Bates J (1988) Econometric issues in stated preference analysis. *J Transp Econ Policy* 22:59–70
- Bendtsen PR (1980) The influence of price of petrol and of cars on the amount of automobile traffic. *Int J Trans Econ* 7:207–213
- Button KJ, Doh S, Hardy MH, Yuan Y, Zhou X (2010) The accuracy of transit system ridership forecasts and capital cost estimates. *Int J Trans Econ* 37:155–168
- Crötte A, Noland RB, Graham DJ (2009) Is the Mexico City metro an inferior good? *Transp Policy* 16:40–45
- DeVany AS (1974) The revealed value of time in air travel. *Rev Econ Stat* 56:77–82
- Flyvbjerg B, Holm M, Buhl SL (2005) How (In)accurate are demand forecasts in public works projects? The case of transportation. *J Am Plann Assoc* 71:131–146
- Friedman M (1953) The methodology of positive economics. In: Friedman M (ed) *Essays in positive economics*. University of Chicago Press, Chicago
- William KM, Mackie PJ (1975) Economics and transportation policy. Allen and Unwin, London
- Hägerstrand T (1969) What about people in regions? *Pap Reg Sci Assoc* 24:7–21
- Heggie I (1978) Putting behaviour into behavioural models of travel choice. *J Oper Res Soc* 29:541–550
- Jones PM (1978) School hour revisions in West Oxfordshire: an exploratory study using HATS. Technical Report, Oxford University Transport Studies Unit
- Kain J (1990) Deception in Dallas: strategic misrepresentation in rail transit promotion and evaluation. *J Am Plann Assoc* 56:184–196
- Kraft K, Domenich TA (1970) Free transit. Heath, Lexington
- Kroes EP, Sheldon RJ (1988) Stated preference methods: an introduction. *J Trans Econ Policy* 22:11–26
- Lago AM, Mayworm P, McEnroe JM (1981) Transit service elasticities – evidence from demonstration and demand models. *J Trans Econ Policy* 15:99–119
- Lindsey R, Verhoef E (2008) Congestion modelling. In: Hensher DA, Button KJ (eds) *Handbook of transportation modelling*, 2nd edn. Elsevier, Amsterdam
- McFadden D (1974) Conditional logit analysis of qualitative choice behaviour. In: Zarembka P (ed) *Frontiers in econometrics*. Academic, New York

- McNally MG, Rindt CR (2008) The activity-based approach. In: Hensher DA, Button KJ (eds) *Handbook of transportation modelling*, 2nd edn. Elsevier, Amsterdam
- Morgan JP (1997) Examining toll road feasibility studies. *Munic Financ J* 18:1–12
- Oum TH, Waters WG, Yong JS (1992) Concepts of price elasticities of transportation demand and recent empirical evidence'. *J Trans Econ Policy* 26:139–154
- Oum TH, Waters WG, Fu X (2008) Transportation demand elasticities. In: Hensher DA, Button KJ (eds) *Handbook of transportation modelling*, 2nd edn. Elsevier, Amsterdam
- Pickrell DH (1992) A desire named streetcar: Fantasy and fact in rail transit planning. *Am Plan Assoc J Am Plan Assoc* 58:158–176
- Reza AM, Spiro HM (1979) The demand for passenger car transportation services and for gasoline. *J Trans Econ Policy* 13:304–319
- Straszheim MR (1978) Airline demand functions on the North Atlantic and their pricing implications. *J Trans Econ Policy* 12:179–195
- Taplin JHE (1980) A coherence approach to estimates of price elasticities in the vacation travel market. *J Trans Econ Policy* 14:19–35
- Wachs M (1990) Ethics and advocacy in foresting for public policy. *Bus Prof Ethics J* 4:141–157

Harvey J. Miller

Contents

37.1	Introduction	706
37.2	Conceptual Foundations of Activity-Based Analysis	706
37.3	Policy and Technology Context for Activity-Based Analysis	708
37.4	Activity Data Collection and Analysis	711
37.5	Activity-Based Modeling	716
37.6	Frontiers in Activity-Based Analysis	720
37.7	Conclusion	722
	References	723

Abstract

Activity-based analysis (ABA) is an approach to understanding transportation, communication, urban, and related social and physical systems using individual actions in space and time as the basis. Although the conceptual foundations, theory, and methodology have a long tradition, until recently an aggregate trip-based approach dominated transportation science and planning. Changes in the business and policy environment for transportation and the increasingly availability of disaggregate mobility data have led to ABA emerging as the dominant approach. This chapter reviews the ABA conceptual foundations and methodologies. ABA techniques include data-driven methods that analyze mobility data directly as well as develop inputs for ABA modeling. ABA models include econometric models, rule-based models and microsimulation/agent-based models. This chapter concludes by identifying major research frontiers in ABA.

H.J. Miller

Department of Geography, University of Utah, Salt Lake City, UT, USA

e-mail: harvey.miller@geog.utah.edu

37.1 Introduction

Activity-based analysis (ABA) refers to treating individual actions in space and time as a basis for understanding human mobility and communication behavior and related systems such as cities, economies, and the physical environment. ABA is replacing aggregate trip-based approaches as the basis for forecasting and knowledge construction in transportation science and urban planning. ABA has long recognized advantages over trip-based approaches, not the least being theoretical validity. In addition, ABA can capture complex constraints and linkages that determine mobility better than aggregate, trip-based approaches. ABA also admits a wider range of policy variables, including non-transportation solutions to mobility problems.

Until recently, the data and computers did not exist to apply ABA to realistic scenarios. These limits have been shattered by increasingly powerful computers but especially by individual-level data available through wireless location-aware technologies embedded in infrastructure, attached to vehicles, and carried by people. These data are enhancing activity data analysis and modeling techniques. They are also leading to a new, data-driven approach to ABA based on exploratory analysis and visualization methods.

The next section of this chapter discusses the conceptual and practical foundations of ABA. It first reviews the traditional, trip-based approach and identifies key weaknesses. The activity-based approach resolves some of these weaknesses by treating mobility and communication not as disembodied flow but as humans conducting the activities that comprise their lives. The [Section 37.3](#) reviews policy and technological changes that are leading to advances and wider application of the ABA approach. [Section 37.4](#) reviews data collection and data analysis methods for ABA. The [Section 37.5](#) discusses activity-based models of travel patterns and urban dynamics using econometric, rule-based, and simulation methods. [Section 37.6](#) identifies ABA research frontiers.

37.2 Conceptual Foundations of Activity-Based Analysis

The past century of transportation science was dominated by a trip-based approach to understanding and predicting human mobility. This approach focuses on isolated acts of mobility as the primary object of study. A *trip* is a movement of a person, goods, and/or vehicle from an origin to a destination (possibly the same location) motivated by positive factors at the locations (push factors at the origin, pull factors at the destination) and attenuated by negative factors related to the cost of mobility between the directed pair. Each trip occurs independently of other activities and trips that occur during individuals' lives. People, events, and activities are atemporal; time is simply a component of mobility cost. Finally, the trip-based approach treats mobile entities not as unique objects but as undifferentiated flows between areas such as traffic analysis zones, postal units, or census geography (although it

can consist of subflows representing different cohorts) (Pinjari and Bhat 2011). Weaknesses of the trip-based approach include (McNally and Rindt 2007):

- No recognition that mobility derives from activity participation
- The treatment of mobility events as resulting from independent and generally unencumbered choice processes, simplifying the complex spatial and temporal constraints that delimit (and sometimes determine) choice
- A focus on utility maximization, neglecting alternate heuristics related to factors such as decision complexity and habits
- A neglect of the roles played by interpersonal relationships and information in influencing activity, mobility, and communication behavior, including *information and communications technologies* (ICTs)

The activity-based approach focuses on the individual and her or his need to participate in activities that have limited availability in time and space. Mobility is not fundamental but an epiphenomenon: it derives from the need to be physically present for many activities and the “inevitability of distance” between activity locations (Ellegård and Svedin 2012). Telepresence via ICTs can substitute for physical presence but can also complement physical mobility by providing more information about events and opportunities as well as capabilities for interpersonal interaction and coordination. Individual and joint allocation of scarce time is the meaningful starting point to understand activity, travel, and communication at all scales: from the tasks required to fulfill daily projects to the annual and decadal dynamics that affect cities, regions, and the planet (Pred 1977). Strengths of the activity-based approach are (McNally and Rindt 2007):

- Recognition that mobility derives from activity participation
- Explicit treatment of the complex temporal and spatial constraints on activity participation and mobility
- Flexibility to accommodate a wide range of decision processes and heuristics
- Explicit treatment of social organization, social networks, and ICTs that influence activity and mobility behavior

Table 37.1 summarizes major components of activity theory. As Table 37.1 illustrates, mobility – trips or tours – is only a component of a more expansive view of human behavior that includes activity patterns and scheduling as well as the social context that influence these activities.

The view that human activities in space and time are the meaningful starting point to understand and manage transportation, cities, and regions dates back to the time-use studies of Chapin (1974) and an influential paper by Jones (1979) that articulated the ABA framework in its contemporary form. But much of the conceptual foundation for ABA was developed by Torsten Hägerstrand in his time geographic framework (Pinjari and Bhat 2011; McNally and Rindt 2007).

Time geography underlies many of the core ideas in ABA, including an ecological perspective on human and physical phenomena, the need to build macro-level explanations from the micro-level and situating travel within a larger context, facilitating the recognition of non-transportation solutions to transportation problems. Basic time geographic concepts such as the individual trading time for

Table 37.1 Elements of activity theory

Element	Description
Activity	The main purpose underlying behavior conducted at a specific location and time interval; often classified as <i>fixed</i> versus <i>flexible</i> activities based on the relative ease of rescheduling and relocation
Activity frequency	The number of times an activity occurs during a given time period
Activity location	Geographic or semantic location where an activity occurs
Activity pattern	Set of activities to be conducted during a specific time interval
Activity schedule	Planned sequence and timing of activities to be conducted during a specific time interval
Time budget	Available time for mobility, communication, and activity participation during a given time interval; often expressed relative to flexible activities and constrained by fixed activities
Trip	Physical movement between activity locations
Interaction	Communication between individuals or locations
Tour	A multi-stop and often multipurpose trip involving several activity locations
Mode	Technique or service used to generate mobility and/or communication behavior
Activity space	Geographic region within which a set of activities occur; can be the composite of discrete activity locations or the smallest spatial region or subnetwork that encompasses the activity locations
Activity environment	Spatiotemporal configuration of activity locations within a given geographic environment
Household	Basic unit of domestic maintenance; influences activity participation through task organization, coordination, and sharing
Social network	Interpersonal relationships, both formal and informal, that influence activity participation
Lifestyle	Socioeconomic and demographic factors that influence activity, mobility, and communication behavior

space in movement among activity locations distributed in time and space may seem trivial since they are so close to everyday life experiences. But this is precisely the point Hägerstrand is making: we neglect seemingly inconsequential but critical factors in our scientific explanations of human behavior; the trip-based approach is an exemplar. Time geography provides a conceptual framework that obligates recognition of basic constraints underlying human existence, as well as an effective notation system for keeping track of these existential facts (Ellegård and Svedin 2012).

37.3 Policy and Technology Context for Activity-Based Analysis

Transportation scientists, engineers, and planners have long recognized the weaknesses of a trip-based approach with respect to validity and accuracy, and the potential of an individual-level, activity-based approach for better understanding and more accurate predictions of transportation and related

human–physical systems. However, until recently there has been little incentive for ABA in policy and planning. There was also little capability with respect to data and computing power.

The last century has witnessed an unprecedented explosion in human mobility due to the development of technologies and services such as steamships, railroads, private automobiles, and commercial aviation. In today's world, people travel to a degree that would have seemed magical to our ancestors. While there are obvious benefits from mobility, there is also increasing recognition of its market failures such as congestion, poor air quality, accidents, sprawled cities, obesity, social exclusion, and global warming. High mobility levels are also increasingly under threat from aging infrastructure that is not being sufficiently renewed, increasing urbanization (especially in the Global South), and increasing motorization as newly emergent economies generate rising levels of wealth.

It is also increasingly difficult to separate mobility and communication behaviors. The telegraph, telephone, and the Internet have revolutionized communication, but these technologies were tightly coupled with location. The rise of mobile telephony and pervasive computing has liberated telecommunication from specific places, allowing it to be more integrated with people and their activities. This is creating tighter, more complex linkages between mobility and communication. Evidence indicates that the “Death of Distance” argument that geographic location would become irrelevant is naïve: communication complements as well as substitutes for mobility, leading to higher mobility demands at all geographic and temporal scales as well as greater complexity of mobility and activity patterns.

Increasing recognition of transportation market failures, threats to mobility, and the tighter integration of mobility and communication behavior have lead to new scientific, policy, and planning initiatives in Europe, North America, and increasingly elsewhere. The business and policy environment for transportation policy and planning is evolving beyond simple measures and prescriptions that focus primarily on measuring throughput relative to cost. There is wider consensus that mobility should be managed, not simply maximized. There is also recognition that evaluating transportation performance requires a fuller range of measures including indicators of effectiveness, equity, community livability, and sustainability. Planners have also realized that solving transportation problems requires thinking outside the system to the broader activity and communication patterns that drive complex mobility behavior. This may include non-transportation remedies for transportation problems (e.g., work flextime, different trading and service hours).

Approaching policy questions from the ABA perspective starts with underlying activity patterns, their interdependencies, and the potential rebound effects that occur from policy changes. Figures 37.1a, b provide a simple example (after Ben-Akiva and Bowman 1998). Figure 37.1a illustrates a daily activity pattern that includes being at home, working, stopping at a day-care center to and from work, and shopping for groceries. Implementing this activity pattern is a single tour from home to the day-care center and work in the morning, shopping in the late afternoon, stopping again at the day-care center and back home in the evening, mostly alone in a private vehicle.

Fig. 37.1 (a) Activity-based approach to policy analysis: before policy intervention.
 (b) Activity-based approach to policy analysis: after policy intervention

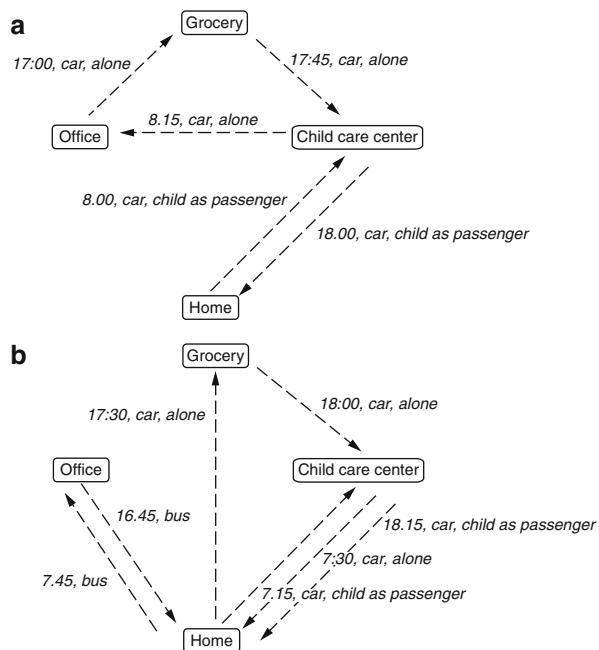


Figure 37.1b illustrates the outcome of a policy intervention: an employer-sponsored public transit incentive combined with higher parking costs. Implementation of the activity pattern now requires three home-based tours: trips to/from childcare center by car in the morning, commuting by bus to and from work during peak times, and shopping by private vehicle in the late afternoon with a stop at the childcare center on the way home. Is this new policy a success? A trip-based approach would likely reach this conclusion since it would focus on the commuting behavior and find a reduction in travel demand by private vehicle. However, an activity-based approach would be more likely to conclude that the new policy was a mixed success due to the shifting of travel and activity patterns and the increase in home-based trips by car. An activity-based approach would capture the linkages between these events and suggest that the transportation policy change should be accompanied by supportive, non-transportation policies such as incentives for day-care centers at work places and/or residential areas.

ABA is more challenging than a trip-based approach: the number of sequencing, timing, location, mode, and route choice possibilities for only a daily activity pattern is combinatorial. There are also a large number of household, social network, and informational linkages that determine daily, weekly, monthly, annual, decadal, and lifetime activity patterns. Activity-based comprehensive urban models also consider the reactions and dynamics of broader infrastructure, economic, sociodemographic, and political systems. Determining a meaningful boundary around the system being analyzed and the level of resolution for representing different components is critical. This requires judgment that considers the scientific

and policy questions being asked, as well as theoretical correctness and consistency (Ben-Akiva and Bowman 1998).

With respect to capabilities for ABA, digital data collection, storage, and processing costs have collapsed to an astonishing degree. *Location-aware technologies* (LATs), digital devices that can report their geographic location densely with respect to time, have become inexpensive and effective. They are increasingly embedded in vehicles and infrastructure and carried by people in consumer products such as smartphones. LATs are generating massive amounts of fine-grained data about mobility and communication dynamics as well as the dynamics of the broader social and environmental systems within which they are embedded. Computers are also much better at handling these data. In addition to dramatic increases in computing power, *geographic information systems* (GIS) and *spatial database management systems* (SDBMS) have evolved well beyond their origins in computer-based paper maps to include a wide range of tools managing, querying, analyzing, and visualizing dynamic and moving objects data. *Social media* available through mobile communication devices allow users to obtain better information transportation systems, share user-generated content, and even participate in management and governance.

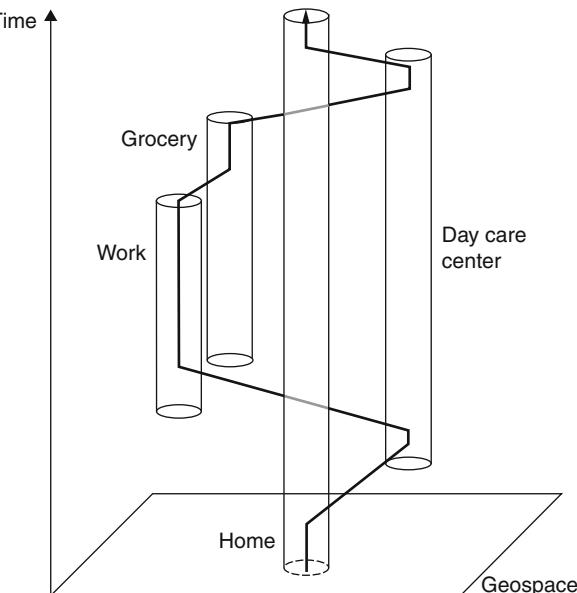
New interdisciplinary fields such as *computational transportation science* (see <http://ctscience.org/>) are emerging to exploit data collection, processing, and communication capabilities to solve vexing and increasingly critical transportation challenges. Private sector companies such as IBM envision smarter transportation, smarter cities, and a smarter and more sustainable planet by collecting fine-grained sensor data, processing these data into meaningful metrics, and sharing this information widely to support more collaborative decision-making (see www.ibm.com/smarterplanet). There are critical privacy questions that must be resolved (discussed below), but these data and tools have the potential to revolutionize transportation science and planning from the “bottom up”: a new science and practice built from individual activities in space and time as the core concept.

37.4 Activity Data Collection and Analysis

ABA includes a rich suite of tools for empirical measurement and analysis of mobility and communication behavior. The conceptual origins for this approach are based in time geography, but this approach has been revolutionized by the rise of LATs and the availability individual-level mobility and communication data. These data can be analyzed directly for empirical patterns. They can also be used as inputs to ABA models, as well as in model calibration and validation. Data-driven methods are also used in *mobility mining*: open-ended exploration of moving objects data to search for novel hypotheses.

Space-Time Paths. The basic conceptual entity in ABA is the fundamental time geographic entity, the space-time path, and its extension, the space-time prism. The *space-time path* represents actual mobility (recorded or simulated) of an entity moving in geospace with respect to time. (A *geospace* is a low-dimensionality

Fig. 37.2 A space–time path among activity locations



space – usually three dimensions or fewer – where distances between location pairs represent shortest path relations in some real-world geography). Figure 37.2 illustrates a space–time path between four activity locations in geographic space (the latter conceptualized as tubes with locations in space and extents in time reflecting their availability).

Semantically, the path is a continuous mapping from time to geospace. In practice, data are typically a sequence of sampled locations strictly ordered by time. Traditionally, these data were collected using recall methods such as travel–activity diaries, prospective methods such as experiments where study participants solve contrived activity and travel scheduling problems. These traditional data collection methods are fraught with problems, including nonparticipation biases, recall biases, and accidental or willful inaccuracies (in the case of travel diaries) as well as difficulties in creating meaningful scenarios (in the case of prospective methods). LATs such as assisted GPS technologies in smartphones allow more accurate and higher volume data collection to support space–time path reconstruction. However, this often comes at the expense of path semantics such as the context for the mobility episode including the planned and executed activities. Semantics can be recovered by overlaying paths with high-resolution georeferenced land-use and infrastructure data. This method can produce errors related to data inaccuracies and activity ambiguities (e.g., what is a person doing while in a coffee house – dining, working, socializing, or some combination of the above?).

The sequence of sample locations can be generated in several ways depending on the data collection method (Andrienko et al. 2008; Ratti et al. 2006):

- *Event-based recording:* Time and location are recorded when a specified event occurs; this is typical of traditional diary methods but also characterizes data

from cell phones, for example, a person calling from a mobile phone generating a location sample.

- *Time-based recording*: Mobile object positions are recorded at regular time intervals; this is typical of GPS and related technologies.
- *Change-based recording*: A record is made when the position of the object is sufficiently different from the previous location; this includes dead-reckoning methods as well as mobile objects database technologies that avoid recording some locations to manage data volume.
- *Location-based recording*: Records are made when the object comes close to specific locations where sensors are located; examples include radiofrequency identification and Bluetooth sensors.

The path must be reconstructed from the temporally ordered sequence of sample locations. The standard method is linear interpolation between temporally adjacent sample points. This requires the least amount of additional assumptions but admits physically unrealistic motions such as infinite acceleration and deceleration at sharp corners. Interpolation via Bezier curves generates a smoother, more physically realistic space–time path (Macedo et al. 2008; Miller 2005a).

Three types of error occur in space–time paths. *Measurement error* refers to error in the recorded location or timestamps. This is equivalent to the well-studied problem of measurement error in polylines in geographic information science. *Sampling error* refers to capturing a continuously moving object using discrete sampling. One way to deal with this is to treat the unobserved segments between sampled locations as an uncertainty region delimiting possible locations for the object between observations (Macedo et al. 2008). This is equivalent to another fundamental time geographic concept, the space–time prism, to be discussed below. *Combined measurement and sampling error* comprises the third type of space–time path error; this is equivalent to measurement error in a space–time prism since under these conditions the space–time prism is a sequence of linked, imperfectly measured space–time prisms.

Space–time paths contain many properties that are useful for understanding human mobility behavior. Analytical methods for paths include (Andrienko et al. 2008; Long and Nelson 2012):

- *Path descriptors* include both *moment-based descriptors* (such as the time, location, direction, and speed at any moment) and *interval-based descriptors* (such as the minimum, maximum, and mean speed; the distribution and sequence of speeds and directions; and the geometric shape of the path over some time interval).
- *Path comparison methods* allow quantitative comparisons among space–time paths, particularly with respect to geometric similarity in space–time and with respect to semantics (such as the sequence of locations visited). Methods include path distance measures such as the Fréchet distance and sequence measures such as least common subsequences.
- *Pattern and cluster methods* for identifying synoptic spatial–temporal patterns from large collections of mobile objects.
- *Individual-group dynamic methods* for characterizing collective movement behavior such as flocking, for example, methods that examine the relative motions among mobile objects.

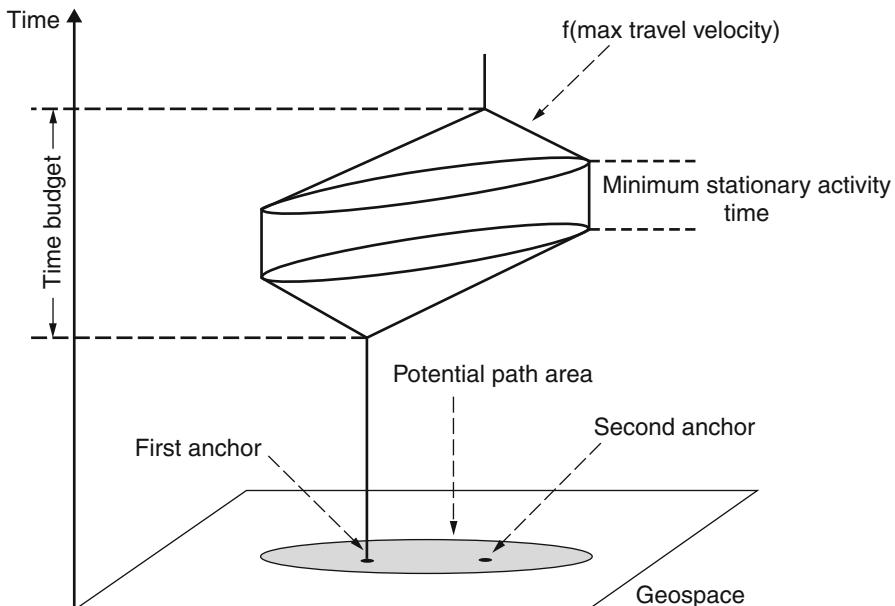


Fig. 37.3 A planar space–time prism

- *Spatial field methods* for translating movement patterns of objects into fields or surfaces that summarize mobility and activity frequency by geographic location.
- *Spatial range methods* for identifying and characterizing the geographic area that contains the observed mobility of one or more mobile objects.

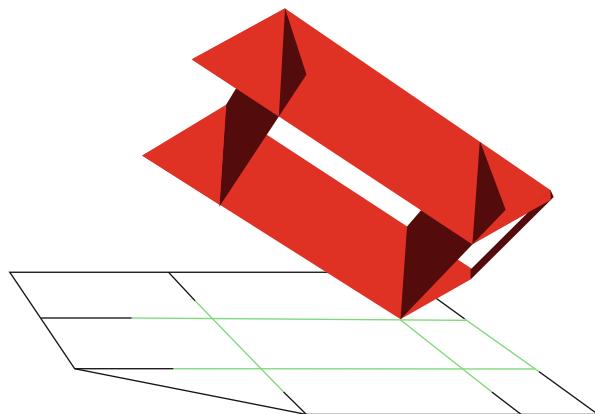
Long and Nelson (2012) provide a succinct but comprehensive review of these methods.

Space–Time Prisms. The *space–time prism* represents potential mobility: it delimits possible locations for a space–time path during some unobserved time interval. Figure 37.3 illustrates a planar space–time prism.

A prism can have two interpretations. As noted above, the prism can be an uncertainty region for an under-sampled space–time path. In contrast, Hägerstrand (1970) conceptualized the prism as a measure of space–time accessibility. The prism encompasses all locations that can be reached during the unobserved time interval given constraints on the object’s speed. Activities, conceptualized as tubes at specific locations with limited extent in time (see Fig. 37.2), must intersect the prism to a sufficient degree (at least as long as the minimal activity time) for the activity to be feasible for that person at that time and location.

The prism is difficult to state analytically over the entire interval of its existence. However, it is tractable to define the prism’s spatial extent at a moment in time as the intersection of only two of three simple spatial regions (Miller 2005a). It is also possible to define space–time prisms within transportation networks (Kuijpers and Othman 2009). Figure 37.4 illustrates a network time prism: the figure illustrates the accessibility locations within the planar network and the corresponding

Fig. 37.4 A network time prism



spatiotemporal region comprising the complete network time prism. In addition to being the envelope for possible space–time paths, these paths also give the prism an internal structure, including unequal visit probabilities within the interior (Winter and Yin 2011).

Prisms contain error propagated from the measured space–time anchors and object speed limits. Error distributions can be numerically generated through Monte Carlo simulation: generate many realizations of the prism and analyze the resulting data. This is a tractable approach for theoretical investigation but is not scalable to practical applications. Alternatively, it is possible to derive analytical characterizations of prisms and prism–prism intersection error in planar space using spatial error propagation theory and implicit function techniques applied to the intersection of circles and ellipses. However, some intersection cases are still open, and it is not scalable beyond pairs of prisms. Required is further investigation into tractable error approximations based on spatial error propagation methods (Kobayashi et al. 2011). More tractable are uncertain network time prisms based on spatiotemporal probability regions (not necessarily connected) for anchor locations and times within the network (Kuijpers et al. 2010).

Prisms can be used as inputs to activity models, in particular choice set or feasible activity set delimitation. Prism-based measures provide vividly different portrayals of accessibility across social, gender, and cultural dimensions relative to traditional place-based measures that tend to mask these differences (Kwan 1998). Prisms can capture activity time constraints within accessibility measures that are consistent with spatial choice, spatial interaction, and consumer surplus theory (Miller 1999).

Path–prism and prism–prism intersections represent potential interaction between two mobile objects. Both can be solved in planar space for a moment in time (Miller 2005a). Scalable techniques also exist for network prism intersections (Kuijpers and Othman 2009). Prism–prism intersections are also useful for capturing the possibility of joint activity behavior in activity-based measures and models (Neutens et al. 2007).

The space–time prism focuses on physical accessibility in geographic space or transportation networks. Path and prism concepts have been extended to encompass interactions within cyberspace (the virtual space implied by networked ICTs). Interaction and accessibility in cyberspace can be treated as direct relationships among space–time paths and prisms (Yu and Shaw 2008) or as indirect relationships mitigated by access to communication technologies (Miller 2005b). It is also possible to treat the STP as existing in a hybrid geo/information space (Coulcelis 2009).

Mobility Mining. Increasing capabilities for collecting and processing mobile objects data is leading to the emergence of *mobility mining* as a new area of research. Mobility mining leverages mobile objects databases with advances in data mining techniques to create a knowledge discovery process centered on the analysis of mobility with explicit reference to geographic context. Mobility mining involves three major phases (Giannotti and Pedreschi 2008):

- *Trajectory reconstruction* from raw mobile objects data. The basic problem was discussed above; the specific problem in this context is to reconstruct trajectories from massive mobile objects data, especially when the data are collected using different methods and sampling methods/rates. This may involve preprocessing steps such as data selection, cleaning, and integrating with other geographic and sociodemographic data.
- *Pattern extraction* involves using spatiotemporal data mining methods to discover interesting (novel, valid, understandable, and useful) patterns in the reconstructed trajectories. Types of patterns include clusters, frequencies, classifications, summary rules, and predictive models.
- *Knowledge delivery* involves verifying and interpreting the discovered patterns, integrating these patterns with background knowledge and communicating this information to support scientific and applied decision-making.

Mobility mining and knowledge discovery from mobile objects databases are hypothesis-generation processes that should lead to more focused and conclusive investigation. These techniques and processes play roles in the scientific process similar to instrumentation such as a telescope, microscope, or supercollider: it allows analysts to see empirical phenomena that would otherwise be obscured or difficult to detect. Empirical patterns discovered during the data mining process are tentative until they have been verified using confirmatory statistics and interpreted in light of background knowledge and theory.

37.5 Activity-Based Modeling

Although theoretically and evidentially suspect, the trip-based approach offers a significant strength, namely, it is relatively straightforward to build scalable comprehensive models of transportation and urban systems that are easily calibrated, verified, summarized, and mapped. It is more challenging to build, verify, and digest comprehensive models built from the micro-level. LAT-based data, geometric growth in computing power, and the hard work of some very smart

people are making activity-based models more realistic, powerful, and understandable. Consequently, ABA is being increasingly applied in policy and planning analysis in Europe, the United States, and other locations.

Depending on the system being modeled, activity-based models can encompass a large number of decision variables over a wide range of temporal and spatial granularities and time frames. In addition, activity-based models are often components in broader comprehensive urban models and linked human–physical process models. Possible components of activity-based models include (Ben-Akiva and Bowman 1998):

- *Activity implementation* involving the execution and possible rescheduling of activity, travel, and communication plans based on empirical conditions in real time. This includes decisions such as mode and route choice, but also fine-grained context-specific behaviors such as speed, acceleration, merging and car-following behavior in automobiles, bicycling behaviors such as obeying stop signs, and pedestrian behavior within crowded environments.
- *Activity scheduling* includes activity selection, activity assignment within household and other social networks, activity scheduling, selection of activity locations, and methods and times for mobility. These events occur frequently and regularly at time scales ranging from real time to hourly, daily, weekly, monthly, seasonally, and annually.
- *Sociodemographic systems* include work, residence, ownership, and other life-altering personal, social, and economic decisions and events such as having children or buying a bicycle. These occur infrequently at the scales from annual to decadal.
- *Urban, social, and economic systems* include the infrastructure, services, institutions, and social and built environments that influence implementation, activity, and lifestyle decisions. These systems operate from real time (e.g., traffic conditions) through annual (e.g., housing dynamics) to decadal and beyond (e.g., compact versus sprawled cities).
- *Physical systems* include material, energy, hydrologic, biological, atmospheric, and other environmental systems that affect and are affected by the other activity domains. These operate in real time (e.g., air quality) to geologic (e.g., climate change).

Activity-based models slice, dice, and combine these components in different ways depending on the modeling domain and scope, as well as the strengths and weaknesses of the particular technique. Major types of activity-based modeling techniques are (i) *econometric models*, (ii) *optimization methods*, (iii) *computational process models*, and (iv) *microsimulation and agent-based models*. Some of these approaches can also be used in combination, for example, econometric models as a component of a larger microsimulation model or a computational process model used to derive agent behavior in an agent-based model.

Econometric Models. Econometric models are among the oldest activity-based modeling strategy, resulting from extending trip-based econometric models to encompass activity choice and trip-chaining behavior. These models have their foundation in the microeconomic theory of consumer choice. They require

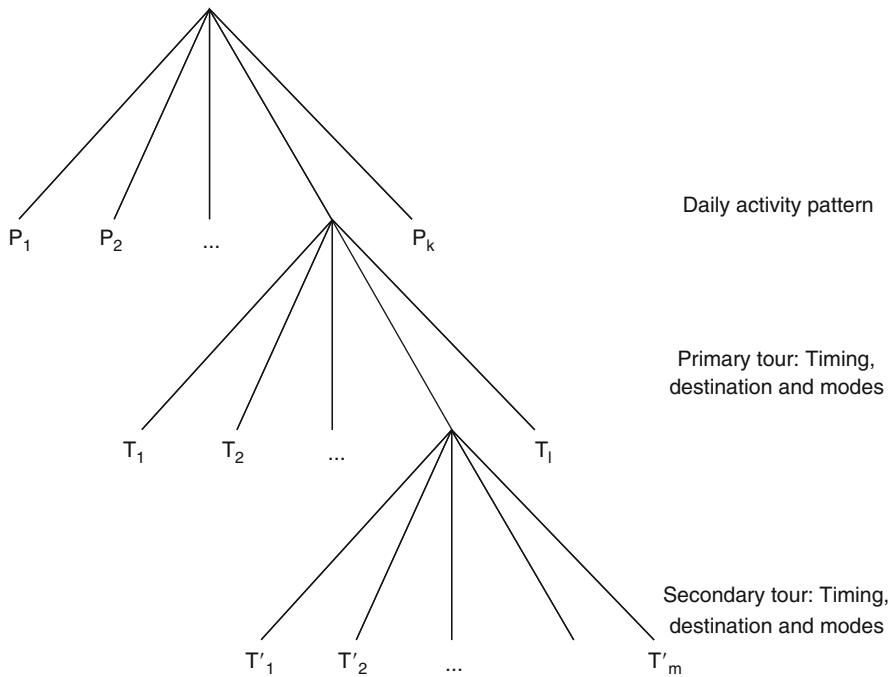


Fig. 37.5 Nested logit representation of activity–travel behavior

specifying relationships between individual attributes, environment factors, and activity–travel decisions in the form of a utility function whose parameters are estimated from empirical data, assuming utility-maximizing choices. Econometric models of activity–travel behavior are often in the form of discrete choice models such as multinomial and nested logit models. Figure 37.5 provides an example of a nested logit representation of activity–travel behavior (after Ben-Akiva and Bowman 1998). Other nesting structures are possible depending on what activity facets are being analyzed. More elaborate econometric structures are also used, such as structural equations, hazard-based duration models, and ordered response models (Ben-Akiva and Bowman 1998; Pinjari and Bhat 2011)

Advantages of econometric models are a rigorous theoretical foundation and mature methodologies for model specification, calibration, and validation. Weaknesses include the empirically suspect assumption that individuals are perfectly rational utility maximizers and the lack of an explicit process theory to describe the activity–travel decision-making (Timmermans et al. 2002).

Optimization Methods. Finding an ideal activity pattern based on criteria such as time, cost, and utility is similar to the problem of finding optimal tours through a transportation network with scheduled pickups and deliveries (Recker 1995). There is a large literature in operations research and management science on problems such as assignment, scheduling, and routing subject to time windows.

These are complex combinatorial problems, but computational search methods have become very sophisticated and powerful. This is a normative approach: the idea is not to replicate real-world behavior but rather generate ideal patterns that can be used as benchmarks for evaluating real-world behavior with respect to efficiency. These comparisons can help identify empirical factors and heuristics that cause people to deviate from ideal patterns.

Rule-Based Models. Computational process models (CPMs) are a system of action-condition pairs (semantically expressed as “if–then” rules) that describe the activity–travel decision process in some empirical domain. Decision rules are often organized according to different subcomponents of the activity system. However, most CPMs focus on activity scheduling and implementation (e.g., Recker et al. 1986). Rules can be derived informally from intuition and knowledge based on previous research. Rules can also be inferred from empirical data using data mining techniques such as decision tree induction and association rules (Arentze et al. 2000).

CPMs are highly flexible, allowing a wide range of heuristics that better represent decision-making in the real world. However, a weakness is the difficulty in enumerating the large number of rules required for even for a modest activity scheduling and implementation problem. CPMs also do not have a mature theory and techniques for testing variables and distinguishing between good and bad models (Buliung and Kanaroglou 2007; McNally and Rindt 2007).

Microsimulation and Agent-Based Models. Microsimulation and agent-based models are computer-based methods for predicting the evolution of a complex system. *Microsimulation* refers to the computer-based modeling phenomena at the disaggregate level to better understand complex dynamics at the aggregate level. Microsimulation has a long tradition in social science, dating back to attempts to modeling the US economy in the 1950s with household and firm behavior as the fundamental units. Microsimulation models tend to fall into two categories. *Static models* typically rely on cross-sectional data and result in no change to the structure of the cross section (e.g., internal composition, sample size) as the model executes over time. *Dynamic models* rely on cross-sectional or longitudinal data and produce changes to the total number of micro-units. Dynamic models are used to forecast and track modifications of entities over longer time periods than static models (Buliung and Kanaroglou 2007).

Agent-based modeling (ABM) is closely related to microsimulation but has a stronger conceptual foundation. ABM views systems as collections of autonomous, adaptive, and interacting agents. An agent is an independent unit that tries to fulfill a set of goals in a dynamic environment. An agent is autonomous if its actions are independent (i.e., makes decisions without an external controlling mechanism) and adaptive if its behavior can improve over time through a learning process. Agents interact by exchanging physical resources and information and/or by reacting to presence or proximity. ABM describes a system from the perspective of its constituent units’ activities; this is appropriate when individual behavior cannot be described adequately through aggregate rules and activities are a more natural way of describing the system than processes (Bonabeau 2002).

The distinction between microsimulation, ABM, and rule-based techniques discussed previously can be vague, particularly in practice. Rule-based methods can be used to drive agent behaviors and microsimulations, and agents can be a central component of broader microsimulation models (e.g., Arentze et al. 2000). It is also possible to link these models with dynamic microscale traffic models to simulate the interrelationships among transportation demand, transportation system performance, and activity scheduling/implementation (see Bekhor et al. 2011).

Advantages of microsimulation and ABM include the explicit representation of micro-level behaviors and processes, the ability to develop and test behavioral theory, better understanding of macro-level processes produced by individual-level behaviors, maintaining the heterogeneity of information (such as individual identity) during simulation, minimization of model bias, better policy sensitivity, integration of processes operating at different temporal scales, and improved model transferability (Bulium and Kanaroglou 2007). Disadvantages include a lack of mature methodologies for calibration and validation, although these models lend themselves to expert engagement and judgment better than traditional, analytical models (Bonabeau 2002). It can also be difficult to make sense of microsimulation models and ABMs; these methods essentially generate a large dataset that must be explored and analyzed. This can be challenging since good scientific practice requires a careful experimental design for parameters that are not empirically derived. The design should vary parameters systematically while holding others fixed to assess the simulation outcomes, often with multiple simulation runs for each parameter combination to eliminate artifacts from random number generators. This can generate a huge amount of simulated results, particularly if there is a large number of parameters and parameter levels to explore.

37.6 Frontiers in Activity-Based Analysis

Much progress has been accomplished in ABA; this progress is likely to continue as favorable policy, computational, and data environments help scientists and practitioners propel it forward intellectually. This section briefly discusses major research frontiers in ABA.

Social Networks. Social networks are at the heart of time geography and ABA: space–time paths bundle to conduct shared activities, prisms intersect to allow this possibility, households are a fundamental unit for activity organization and sharing, and activity coordination and adjustments cascade through broader activity and social systems. Time geography and ABA are an ecological approach to transportation, cities, and societies with a complex web of interconnections (Pred 1977; Ellegård and Svedin 2012). Capturing the social network influences on activity, mobility, and communication behavior is a very active frontier in ABA (Neutens et al. 2008).

A major challenge in capturing social networks in ABA concerns basic definition, measurement, and data collection. Social networks can range from a few intimate individuals to hundreds of Facebook friends. The problem is that all of

these networks are relevant to activity behavior depending on the context. Measuring social networks is also difficult, particularly more genuine and enduring networks. Social influence within these networks can also vary depending on formal and informal relations. Finally, social networks have complex topologies such as Small World configurations that can generate complex dynamics.

LATs and social media can inform social networks in ABA. As mentioned above, path–path, path–prism, and prism–prism relationships indicate the possibility of social interaction, and methods for collective mobile objects data analysis are improving. Problems include dealing with coincidental proximity (e.g., friends versus strangers in a coffeehouse) and activity ambiguity (e.g., a coffeehouse again). Location data error is also a challenge: this can be substantial for some LATs in some environments (e.g., GPS receivers in city centers, cellular network location in rural areas).

Social media are convincing millions of people to share details of their lives online. The implications of these data for understanding and predicting activity, travel, and communication behavior should be obvious, including that people use these media to plan and coordinate activities. Challenges include nonrepresentation biases and unstructured data. Social media participants are not scientifically sampled, nor do people share everything about their lives (with some notable exceptions). Nevertheless, the massive size of these databases makes them valuable. Social media data are also unstructured: nonquantitative data such as text and imagery. Intriguingly, these data are increasing georeferenced due to social media applications in smartphones. One way to treat these data is from a mobility mining perspective: use social media data to generate hypotheses that can be tested with more focused, confirmatory techniques and scientifically sampled or experimental data.

Unfortunately, access to LAT and social data can be circumscribed due to proprietary and competitive reasons. This has the danger of leading to a computational approach that will revolutionize the social sciences but only as practiced in private sector companies and secret government agencies (Lazer et al. 2009).

Big Data and Knowledge Delivery. Big Data refers to data that has *high volume* (massive databases with millions of records), *high variety* (structured and unstructured data), and *high velocity* (often in real time). The Big Data mantra is to keep all of these data since they may be useful; the astonishing collapse in data storage costs over the past two decades makes this possible. In many locations in the world, we are moving toward sensed transportation systems with sensors embedded in infrastructure and vehicles, as well as high-resolution but remotely positioned sensors such as LiDAR. These data combined with consumer LAT data and social media will generate orders of magnitude more data about transportation and cities than currently exist.

A previous section of this chapter discussed the role of mobility mining in ABA. Research frontiers include not only dealing with massive transportation, mobility, and communication data but delivering actionable knowledge to decision-makers sufficiently fast, so they can act before the knowledge is irrelevant. This is a challenging frontier that involves elements of exploratory and confirmatory analysis as well as decision support.

Big Data also has the potential to create more collaborative transportation and social systems. This is a major motivation behind IBM's Smarter Planet initiatives. Collaborative transportation systems can range from ride/vehicle sharing to long-term strategic decision-making about transportation and urban futures. The challenge is to create not only the knowledge delivery techniques discussed in the previous paragraph but also the tools and environments for sharing, collaboration, and collective governance.

Locational Privacy. The benefits of an ABA reinvigorated through more data and computational power may not be realized if there is a public backlash due to abuses of these data. *Locational privacy* is the concept that the space–time signature that comprises activity patterns can reveal much about a person and her/his activities. This is a fundamental change: as the United States Supreme Court commented during a recent decision, LATs provide not isolated facets but a person's entire life.

Locational privacy protection strategies include *regulation*, *privacy policies*, *anonymity*, and *obfuscation*. Regulation and privacy policies define unacceptable uses of location data. Anonymity detaches locational data from an individual's identity. Obfuscation techniques degrade locational data through deliberate undersampling, aggregation, introducing measurement error, or some combination of the above. Scientific challenges include new research ethical protocols for dealing with location data, especially user-generated content and remote but high-resolution sensors that can reveal things and activities that were previously considered private. Another scientific challenge is dealing with deliberately degraded locational data; spatial and spatiotemporal error methods for mobile objects data are still lacking to a large degree. More generally, societies need to have conversations about the acceptable and unacceptable uses of these data if their role in building better transportation systems and communities is to continue its remarkable progress.

37.7 Conclusion

Activity-based analysis (ABA) is emerging as the dominant approach in transportation science and planning (Timmermans et al. 2002). It is a theoretically sound approach to transportation, cities, societies, and human–physical systems that focuses on a person's activities in time and space as the foundation. Changes in policy are encouraging a wider view of transportation, and the increasing availability of individual mobility data and scientific advances inspired by this favorable environment are making ABA methods scalable to realistic scenarios and problems.

Data-driven methods allow high-resolution measurement of fundamental ABA entities such as the space–time path (representing actual mobility) and the space–time prism (representing potential mobility, interpreted as path sampling error or space–time accessibility). There is a wide range of methods for measuring, comparing, and summarizing collections of space–time paths, but fewer methods for the space–time prism. These data can be used for empirical investigation, mobility data mining, and as inputs to ABA modeling.

ABA models attempt to solve or simulate activity behavior. Most models focus on the activity scheduling and implementation problems. These ABA core models can be linked with transportation system performance models to capture the dynamics of mobility demand and system response. These core models can also be embedded in broader models of cities, sociodemographics, and physical systems such as airsheds. Major modeling approaches include econometric models, optimization methods, computational process models, and microsimulation/agent-based models.

There are several ABA research frontiers; these include social networks, delivering knowledge in the face of Big Data, and location privacy. Progress along these frontiers will support the continuing rise of ABA in understanding and planning transportation and related systems.

Acknowledgments Dr. Waled Othman (University of Zurich) provided the Mathematica code to generate the network time prism (Fig. 37.4); this is available at <http://othmanw.submanifold.be/>. Ying Song (University of Utah) generated some of the graphics. Ying Song and Calvin Tribby (University of Utah) provided valuable comments on this chapter.

References

- Andrienko N, Andrienko G, Pelekis N, Spaccapietra S (2008) Basic concepts of movement data. In: Giannotti F, Pedreschi D (eds) Mobility, data mining and privacy. Springer, Heidelberg, pp 15–38
- Arentze T, Hofman F, van Mourik H, Timmermans H (2000) ALBATROSS: multiagent, rule-based model of activity pattern decisions. *Transp Res Rec* 1706:136–144
- Bekhor S, Dobler C, Axhausen K (2011) Integration of activity-based and agent-based models: case of Tel Aviv, Israel. *Transp Res Rec* 2255:38–47
- Ben-Akiva M, Bowman JL (1998) Activity based travel demand model systems. In: Marcotte P, Nguyen S (eds) Equilibrium and advanced transportation models. Kluwer, Boston, pp 27–46
- Bonabeau E (2002) Agent-based modeling: methods and techniques for simulating human systems. *Proc Natl Acad Sci* 99(suppl 3):7280–7287
- Buliung RN, Kanaroglou PS (2007) Activity–travel behaviour research: conceptual issues, state of the art, and emerging perspectives on behavioural analysis and simulation modeling. *Transp Rev* 27(2):151–187
- Chapin FS (1974) Human activity patterns in the city: things people do in time and space. Wiley, London
- Couclelis H (2009) Rethinking time geography in the information age. *Environ Plan A* 41(7):1556–1575
- Ellegård K, Svedin U (2012) Torsten Hägerstrand's time-geography as the cradle of the activity approach in transport geography. *J Transp Geogr* 23:17–25
- Giannotti F, Pedreschi D (2008) Mobility, data mining and privacy: a vision of convergence. In: Giannotti F, Pedreschi D (eds) Mobility, data mining and privacy. Springer, Heidelberg, pp 1–11
- Hägerstrand T (1970) “What about people in Regional Science?” *Papers of the Regional Science Association* 24(1):6–21
- Jones PM (1979) New approaches to understanding travel behaviour: the human activity approach. In: Hensher DA, Stopher PR (eds) Behavioral travel modeling. Croom-Helm, London, pp 55–80

- Kobayashi T, Miller HJ, Othman W (2011) Analytical methods for error propagation in planar space-time prisms. *J Geogr Syst* 13(4):327–354
- Kuijpers B, Othman W (2009) Modeling uncertainty of moving objects on road networks via space-time prisms. *Int J Geogr Inform Sci* 23(9):1095–1117
- Kuijpers B, Miller HJ, Neutens T, Othman W (2010) Anchor uncertainty and space-time prisms on road networks. *Int J Geogr Inform Sci* 24(10):1223–1248
- Kwan M-P (1998) Space–time and integral measures of individual accessibility: a comparative analysis using a point-based framework. *Geogr Anal* 30(3):191–216
- Lazer D, Pentland A, Adamic L, Aral S, Barabási A-L, Brewer D, Christakis N, Contractor N, Fowler J, Gutmann M, Jebara T, King G, Macy M, Roy D, Van Alstyne M (2009) Life in the network: the coming age of computational social science. *Science* 323(5915):721–723
- Long JA, Nelson TA (2012) A review of quantitative methods for movement data. *Int J Geogr Inform Sci* (in press)
- Macedo J, Vangenot C, Othman W, Pelekis N, Frentzos E, Kuijpers B, Ntoutsi I, Spaccapietra S, Theodoridis Y (2008) Trajectory data models. In: Giannotti F, Pedreschi D (eds) *Mobility, data mining and privacy*. Springer, Heidelberg, pp 123–150
- McNally MG, Rindt CR (2007) “The activity-based approach”, working paper UCI-ITS-AS-WP-07-1, Institute of Transportation Studies, University of California-Irvine
- Miller HJ (1999) Measuring space-time accessibility benefits within transportation networks: basic theory and computational methods. *Geogr Anal* 31(2):187–212
- Miller HJ (2005a) A measurement theory for time geography. *Geogr Anal* 37(1):17–45
- Miller HJ (2005b) Necessary space-time conditions for human interaction. *Environ Plan B Plan Design* 32:381–401
- Neutens T, Witlox F, van de Weghe N, DeMaeyer P (2007) Space-time opportunities for multiple agents: a constraint-based approach. *International Journal of Geographic Information Science* 21(10):1061–1076
- Neutens T, Schwanen T, Witlox F, De Maeyer P (2008) “My space or your space? Towards a measure of joint accessibility”, computers. *Environ Urban Syst* 32(5):331–342
- Pinjari AR, Bhat CR (2011) Activity-based travel demand analysis. In: de Palma A, Lindsey R, Quinet E, Vickerman R (eds) *Handbook in transport economics*. Edward Elgar, Cheltenham, pp 213–248
- Pred A (1977) The choreography of existence: comments on Hägerstrand’s time-geography and its usefulness. *Econ Geogr* 53(2):207–221
- Ratti C, Pulselli RM, Williams S, Frenchman D (2006) Mobile landscapes: using location data from cell phones for urban analysis. *Environment and Planning B* 33(5):727–748
- Recker WW (1995) The household activity pattern problem: general formulation and solution. *Transp Res* 29(1):61–77
- Recker WW, McNally MG, Root GS (1986) A model of complex travel behavior: part i. Theoretical development. *Transp Res Part A* 20(4):307–318
- Timmermans HJP, Arenze T, Joh C-H (2002) Analyzing space-time behavior: new approaches to old problems. *Prog Hum Geogr* 26(2):175–190
- Winter S, Yin Z-C (2011) The elements of probabilistic time geography. *Geoinformatica* 15(3):417–434
- Yu H, Shaw S-L (2008) Exploring potential human activities in physical and virtual spaces: a spatio-temporal GIS approach. *Int J Geogr Inform Sci* 22(4):409–430

Nigel Waters

Contents

38.1	Introduction	726
38.2	The Origins of Network Science in Regional Science	727
38.3	The Dark Ages of Network Science and the Resurgence of the 1990s	727
38.4	Network Science at the End of the Millennium: New Findings	729
38.5	The New Importance of Social Media	731
38.6	The Development of Explicit Social Network Platforms	735
38.7	Measuring Individual Influence Within Social Networks	736
38.8	Recent Developments in Network Science	736
38.9	The Decline of Distance and the Need for a Second Law of Geography	737
38.10	Conclusions	738
	References	739

Abstract

This chapter begins with a discussion of how communications technologies have reduced the influence of distance on the location of economic activity. The origins of network analysis in regional science are described. The importance of social networks and social network science in sociology and related disciplines during the 1970s, 1980s, and 1990s is explained. This is followed by a discussion of new discoveries concerning the structure of the Internet that took place in the late 1990s. The rise of social media, the continued development of social network science, and the popularity of social network sites such as Facebook, Twitter, and LinkedIn in the new millennium are then depicted

“I almost wept when I awoke, and found that you had appeared to me in Time and not in Space, alas!” Thomas Hardy, *The Woodlanders*.

N. Waters

Department of Geography and Geoinformation Science, George Mason University,
Fairfax, VA, USA
e-mail: nwaters@gmu.edu

along with the most recent research findings that derive from connectivity and contagion processes within social networks. The chapter concludes with an account of methods for determining the importance of distance in influencing social and economic activity in the new world of social networks.

Abbreviations

RS	Regional science
SM	Social media
SN	Social network
SNA	Social network analysis
UGC	User-generated content

38.1 Introduction

Today, there can be little doubt that much economic activity is taking place on and because of social networks (SNs). However, the more interesting questions that we will consider in this chapter are the following: What aspects of SNs influence these activities, and do SNs affect previously established patterns of spatial activities and spatial interaction? Do they reflect and support existing geographical patterns, or do they create new geographies due to a weakening of the influence of distance?

From its earliest days, regional science (RS) was seen as an interdisciplinary activity that privileged a spatial and analytical approach to the social sciences, emphasizing location theory and spatial economics. The influence of physical, geographical distance and its impact on transportation costs and, more specifically, the location of economic activity were developed from the earlier work of many of the *German theorists* of the nineteenth and early twentieth centuries including von Thuenen, Alfred Weber, Walter Christaller, and August Loesch (for a review of the work of each of these authors, see the classic economic geography text by Smith et al. 1969). Interest in networks was confined to physical transportation network canals, roads, railroads, and subsequently airline networks. Always distance and the physical separation of economic activities lay at the core of location theory.

Beginning in the 1990s, a trio of mutually supportive technologies, namely, the Internet, social media, and social networks, began to weaken the influence of distance leading to such phenomena as globalization and the increased integration of national economies through the expansion of multinational corporations. The influence of these and earlier technologies, including the telex and telephone, was discussed in popular texts such as *The Death of Distance* (Cairncross 2001). Cairncross clearly overstated the case by arguing that distance no longer mattered in terms of the location of economic activity and the strength of social interaction between locations but that it mattered less was beyond debate. Other researchers (Rietveld and Vickerman 2004), at least in the case of the influence of transport costs in regional science, have demurred from this view, suggesting that this supposed death of distance is “premature.” The question as to which view is most appropriate in the case of SNs remains ambiguous.

In this chapter on SNs and regional science, we will consider the development of a network science, the development of social media, the rise of social networks such as Facebook, the development of social media as a marketing tool, and the current state-of-the-art in determining just how important distance remains in a regional science that makes use of social network databases.

38.2 The Origins of Network Science in Regional Science

In the 1950s and 1960s, geographers and regional scientists were well aware of each other's research. Indeed the seminal works of von Thuenen, Weber, Christaller, and Loesch, mentioned above, were equally prominent in the research of both North American and British regional scientists and geographers. This work was subsequently codified and organized into a coherent argument in Haggett and Chorley's (1969) text: *Network Analysis in Geography*. Haggett and Chorley's summary and the research on which it was based emphasized the topological properties of networks and in particular both the structure of the network as a whole and the importance of nodes within the network as expressed by the number of their immediate connections in the network and their centrality within the network.

While this text was well received and indeed with its explicit discussion of *graph theory* laid the foundation for a network science, little further work on the topological structure of transportation networks occurred in the ensuing years, and for many regional scientists, this line of research appeared to have reached a natural conclusion with little promise of further insights (Waters 2006).

38.3 The Dark Ages of Network Science and the Resurgence of the 1990s

Regional scientists and geographers paid scant attention to network science in the decades that ensued, that is, throughout the remainder of the 1970s, the 1980s, and into the 1990s. This was not true of all the social science disciplines and sociology, especially, saw the importance of continued research in the analysis of the topological properties of social networks. During this time, there was a rapid, exponential rise in Sociological Abstracts that used the term "social network" in their abstract or title between 1970 and 2000. Important summaries of the state-of-the-art of social network analysis (SNA) in the social and behavioral sciences were provided by Scott (1992) and by Wasserman and Faust (1994). The latter authors included a review of the SN literature from the 1950s to the 1990s, a discussion of the methodologies and mathematics of SNA, and an appendix of software packages.

Scott (1992) provided a relatively complete synopsis of the subject of SNA within the discipline of sociology, noting that SNA had been used in various sociological studies including network studies of the financial powers among bank directors, social mobility, kinship and class structures, contacts in gangs

and other outlaw societies, and even science citations. Much of this work evolved from the research of *Harrison White* and his graduate students in the Department of Social Relations at Harvard University in the 1960s and early 1970s. These students included Stanley Milgram and Mark Granovetter, two of the most influential contributors to the SN literature. In 1969 Milgram and a colleague published a seminal article on *The Small World Problem* (Travers and Milgram 1969). It was here that he popularized the notion that everyone was connected by an average of “six degrees of separation,” an idea that was to have a major impact on subsequent SN research (see below). Granovetter’s original paper on *The Strength of Weak Ties* (Granovetter 1973) first identified the importance of “bridges” between tightly knit clusters within social networks, and this also became one of the most widely cited references in the SN research literature.

White’s students spread out across North America, accepting appointments in leading universities and establishing productive SN research centers. A journal, *Social Networks*, and an influential, peer-reviewed newsletter, *Connections*, published by the International Network for Social Network Analysis, were quickly established.

Scott’s original handbook (1992) provided a complete guide to SNA describing the history of the subdiscipline in sociology and then the representation of SNs as graphs or *sociograms* with links or ties to the nodes or points that represented the individuals that were connected with each other. Scott also gave detailed descriptions of methods for the storage of SN data and of measures of centrality of individuals within the network and the importance of nodes (individuals) that linked together network clusters (Granovetter’s bridges).

Scott (1992) reviewed existing software packages for conducting SNA including *UCINET* from the University of California, Irvine, and the *PAJEK* software that was specifically designed to handle large data sets such as those that were then beginning to emerge on the Internet. UCINET and PAJEK have remained two of the most popular software packages for SNA and SN visualization. Recent lists of SNA software may be found in the many new texts that are constantly appearing in this widely researched field (e.g., Scott’s new handbook (Scott, 2011) and references cited below). It should be noted that many of the analytical procedures discussed in the texts by Scott and by Wasserman and Faust were minor developments of earlier procedures. These included measures of nodal importance such as the degree of a node (namely, the number of ties attached to it) or measures of centrality (i.e., variations of measures designed to assess a node’s position within the network’s topological structure). Other analytical procedures were extensions of multivariate techniques such as cluster analysis, principal component analysis, and multidimensional scaling that had been applied to networks in the 1960s and 1970s (Waters 2006). What was new was the ability of these software packages to collect and store the enormously large, Internet-based data sets and to visualize the structure of these networks in an informative and intuitively pleasing manner.

38.4 Network Science at the End of the Millennium: New Findings

Toward the end of the 1990s, a series of papers (Watts and Strogatz 1998; Barabasi and Albert 1999) showed that many networks that exist in the natural world and also in socially constructed environments such as the World Wide Web exhibited the so-called *small world phenomenon* demonstrated in Milgram's experiment three decades earlier. Networks with large numbers of nodes were found to have surprisingly small average path lengths due to the fact that a small percentage of nodes had a large number of connections, that is, a high “*degree*” *number*. These nodes acted as hubs or bridges providing shortcuts across the topological structure of the network. The degree frequency distribution was shown to follow a *power law* or *Pareto distribution* with a so-called *heavy tail*. Barabasi and Albert (1999) referred to such SNs as being *scale-free* or *scale invariant*, and they argued that this arose due to *preferential attachment* where the probability of attaching to an existing node was proportional to its degree.

Pareto distributions and the *rank size rule* for city size distributions have long been observed to be of great significance in the physical world of regional science with new mechanisms for the emergence of power laws in urban systems and elsewhere being suggested by Reed (2001). However, it should be noted that more recently Willinger et al. (2009) have challenged what has been described as the “*scale-free Internet myth*.” The scale-free myth has resulted in observations that such networks are robust to random failures because these are likely to occur at nodes with low-degree connectivity (because they are more common) and sensitive to targeted attack, such as terrorist activity, because such attacks would focus on the high-degree hubs. Willinger et al. (2009) argued that much of the work used to establish the scale-free myth was flawed in its collection and sampling design and then demonstrated that if the construction of an Internet router network is conceived as a constrained optimization problem in which traffic is distributed as a gravity model (well known to regional scientists), then preferential attachment models become irrelevant. High degree variance is simply a result of high variance in demand for bandwidth. Adding to the debate, Strogatz (2005) notes that there may be so many paths to the realization of a power law scaling that in the absence of other explanations these observations may simply be “all sound and fury” and signify “nothing” or alternatively, as in the case of city size distributions, it is often the deviations from the model, such as primate cities, that are truly interesting.

Markoff and Sengupta (2011) describe a recent study of 721 million Facebook users where the average degree of separation was 4.74, but when the results were restricted to the USA, the average separation dropped to 4.37. This research is important in regional economics because it emphasizes the global reach of SNs, the density of connections, and the linkage of clusters that allows for those engaged in Internet commerce to extend their reach with minimum marketing costs. Furthermore, the short paths linking everyone to everyone else in these very “small worlds” may well explain why ideas and products can go “viral” with such rapidity and why it is attractive for those engaged in both enforcement and insurrection alike to use the power of SNs for their own goals.

A mechanism for the ever-decreasing number of ties or links from one member of an SN to another has been suggested by the experiments of Watts et al. (2002). They suggest that SNs are becoming increasingly “searchable,” that is to say individual users are becoming more and more able to direct messages through networks to targeted individuals. If this is true, then it is also likely to be even more feasible for businesses also to take advantage of the inherent characteristics of SNs for economic gain. For this to be the case, the SNs used in the experiments designed by Watts et al. were endowed with the following characteristics that would appear to be plausible for real-world SNs. First, individual users had specific characteristics that allowed them to form groups; second, these groups were hierarchically organized; third, the groups themselves were the basis for social interaction; fourth, individuals were hierarchically clustered in more than one dimension, for example, by occupation and geography, and these dimensions were independent (perhaps not altogether a realistic assumption); fifth, individuals constructed a measure of *social distance* between themselves and others based on their perceived similarity over all dimensions (a global pseudo or structural distance); and sixth, individuals forwarded a message only to their direct connections within the network. For the experiment, parameter choices were made consistent with the inferred SNs in Milgram’s original experiment. The model, which is applicable to all peer-to-peer communication systems, was shown to yield results that are statistically indistinguishable from those of Milgram, though whether this guarantees that they are similar in other respects is open to question.

Connections, represented by SN ties, are important because they determine the structure of the network. Also the characteristics of ties are variable. They may be professional or social or they may be permanent or ephemeral. Equally important is *network contagion* or sharing because this relates to what passes across a network – information, ideas, money, product (digital or otherwise), disease, life-style, and happiness, among many others. According to Christakis and Fowler (2011), connection and contagion are governed by the following five rules.

Rule one relates to the fact that SN users shape and, indeed, constantly reshape their network. Usually we connect to others that are like us in terms of socioeconomic characteristics such as income, education, ethnicity, and language, among others. This is known as *homophily*. Socially, these others are likely to live nearby and thus be highly spatially autocorrelated reinforcing existing patterns of economic activity. Professionally, this is less likely to be so and this will have novel impacts on economic activity, enhancing, for example, tendencies toward globalization. SN users have considerable choice in terms of how they structure their connections including how many people they connect to and how dense are their connections and, to some degree, their centrality within their network. It is also important to note that some individuals will connect late or not at all. These are the “laggards” and nonparticipants, and it is just as important to be aware of their attributes and characteristics. Rule two states that our place in the network, our immediate connections and degree of centrality, affects our social and economic behavior. Rule three is closely associated for it states that our friends, our immediate connections, affect us. If they are happy, then we are likely to be happy too;

if they are obese, then we will have that tendency as well. Not only that but rule four states that this influence extends also to our friends' friends and even more surprisingly to our *friends' friends' friends* (three degrees of separation), after which the effect peters out. Christakis and Fowler (2011) have documented the impact of these rules in a series of pioneering studies explaining their effect on the spread of obesity, the spread of happiness, and the dynamics of smoking. Rule five relates to the emergent properties of the network itself. Thus, the network, which in these cases is referred to as an *excitable medium*, may develop properties that none of its members is aware of initially. This may occur with flash mobs, insurgent activities, or in disaster management.

38.5 The New Importance of Social Media

For our purposes, we define social media (SM) as the all-encompassing concept within which social networks are included. Kaplan and Haenlein (2010) provide a brief history of the development of social media and a definition that distinguishes social media from the related concepts of *Web 2.0*, *user-generated content*, and social networks. According to Kaplan and Haenlein, the history of social media has its earliest origins in 1979 with the development of Usenet by Tom Truscott and Jim Ellis. This system allowed for the posting of messages and thus a "society" of users, but although popular and still in use, it was not until almost 20 years later that the SM era really took off with the development of Open Diary by Bruce and Susan Abelson. This was followed by what was initially known as web logging which was almost immediately abbreviated to blogging. Eventually these activities were to spawn social networking sites such as MySpace and Facebook, founded in 2003 and 2004, respectively. These types of SM activities relied on both the development of Web 2.0 and the growing popularity of user-generated content (UGC).

Web 2.0, a term first used in 2004, provides the technical and ideological basis for SM activities. Web 2.0 was facilitated by the development of new technologies such as Adobe Flash, RSS (Real Simple Syndication), and AJAX. Respectively, these allowed for the addition of animation and video, for web feeds to update content rapidly, and for the continuous updating of websites without affecting the display. Through the use of these technologies, Web 1.0 evolved from a platform for individually created content into Web 2.0 where content was generated in a collaborative fashion, Wikipedia being perhaps the iconic example. UGC, a term that came into popular use in 2005, represents all the various ways in which people make use of SM. Kaplan and Haenlein (2010), p.61 use both Web 2.0 and UGC to define SM in the following way: "Social Media is a group of Internet-based applications that build on the ideological and technological foundations of Web 2.0, and that allow the creation and exchange of User Generated Content."

Kaplan and Haenlein (2010) employed ideas from media research to produce a two-factor classification of SM types. The first factor concerns social presence and media richness and is based on the concept of social presence theory which

Table 38.1 Social media classification (adapted from Kaplan and Haenlein 2010)

		Social presence/media richness		
		Low	Medium	High
Self-presentation/ self-disclosure	Low	Collaborative projects (e.g., Wikimapia, Wikipedia)	Content communities (e.g., YouTube, Flickr, Digg, TravBuddy)	Virtual game worlds (e.g., World of Warcraft, Runes of Magic)
	High	Blogs	Social networks (e.g., MySpace, Facebook, LinkedIn)	Virtual social worlds (e.g., Second Life; SmallWorlds)

measures the degree of physical, visual, and acoustic interaction that is achieved between the individuals communicating. Media richness is closely related to the concept of social presence. The second factor is based on social processes and concerns both self-disclosure and self-presentation. A selection of social media examples using this categorization is shown in [Table 38.1](#). This chapter is primarily concerned with the high self-disclosure/self-presentation and medium social presence/media richness category that is appropriately labeled as “social networks.”

The history of social media may be traced back to 1997 (Boyd and Ellison [2007](#)) when the Six Degrees SN site allowed users to create profiles, designate friends, and incorporate friends of friends to their lists. To navigate what they describe as the SM jungle, Kietzmann et al. ([2011](#)) offer an “ecology” based on seven building blocks of SM functionality, and for each they explore the implications of this functionality. The seven building blocks are, respectively, identity, conversations, sharing, presence, relationships, representation, and groups. Here, following the discussion in Kietzmann et al. ([2011](#)), we treat each attribute in turn, noting that these seven characteristics are by no means mutually exclusive.

Identity concerns the extent to which users reveal their personal characteristics. Usually these include their names and the standard socioeconomic parameters such as age, gender, profession, and location but not education and income (although owners of SN sites might infer the latter from correlations with the other characteristics). The entire package of socioeconomic indicators can then be used to create a geodemographic profile that can be used for subsequent “target marketing.”

SM sites usually provide various levels of data privacy controls for the protection of their members along with various filters and shields that also protect against information overload. Different sites may produce quite different identities for the same user. Thus, a LinkedIn “professional” profile may be quite different from a Facebook “social” profile. Secondary services such as DandyID allow SN users to record these different profiles in a single location. Dandy ID provides social analytics tools for allowing their users to determine how people are engaging their clients across the entire SM spectrum. For SM users that want to provide access to their profiles without revealing their identity, sites such as OAuth provide the necessary protection tools.

The conversations attribute reflects the extent to which SM site users communicate with each other. They may do this for personal reasons, for advocacy, or for

commercial purposes. These conversations may be brief as in the case of *microblogs* such as Twitter, Jaiku, Plurk, and Tumblr (Kaplan and Haenlein 2011) or more extensive as in traditional blogs. Microblogs can create what has become known as *ambient awareness* information that relates to the immediate surroundings as defined in time or space or both. Such information can be enormously useful for political action and location-based services or to assist in the marketing process. In a special issue of the journal Business Horizons dedicated to social media, Kaplan and Haenlein (2011) note that microblogs are useful in all three phases of marketing, namely, the prepurchase, purchase, and post-purchase phases. These include marketing research, marketing communications, and customer service, respectively. Examples cited by these authors include Dell Computer's Communities and Conversations team which has used customer tweets to redesign their Inspiron Mini 10 computer (marketing research), the airline JetBlue that uses its daily “cheep” tweets to fill empty seats (marketing communications), and Whole Foods Market to manage customer complaints (customer service).

Sharing on an SM site relates to the degree to which users receive or distribute and exchange content. Different SM websites are focused on sharing different objects of sociality. For example, LinkedIn users share data relating to professional careers, Flickr pictures, and YouTube videos. Important issues are how to grow the type of media that is shared and how to manage the shared content that might violate copyright laws or be offensive or inappropriate. Growing the media that are shared can be achieved by acquiring start-ups that offer new services. In 2012 Facebook bought both Lightbox, a photo-sharing site, and Karma, an app for the new activity of “social gifting.” Sharing is commonly associated with social shopping services or deal-of-the-day sites such as Groupon, LivingSocial, and Google Offers, among others.

In this chapter, we are concerned with the impact of SNs on economic geography and regional science. A primary impact, as we note throughout this chapter, is the weakening of geography and distance, albeit that this impact varies with the type and purpose of the SN site. A second impact is that SNs allow for the creation of alternative, informal economies. Products can be bartered with no exchange of funds. This can be facilitated by the SN whereby contacts are made but products are exchanged in local markets or trade can take place entirely on the SN site through sharing. SN systems based on sharing are flourishing in countries such as Greece where the recession has caused a freezing of credit and record high unemployment making these systems a more palatable alternative. Systems can be operated using complementary or community currencies such as Cyclos, LETS (local exchange trading systems), and time banking, where, in the latter instance, time is used as a unit of currency. In Greece, the alternative local currency (ALC) that has replaced the Euro in some markets in towns such as Volos is referred to by its Greek acronym, TEM.

Presence relates to the ease with which users can determine if another user is available. As an example, Skype includes the following levels of accessibility: online, away, do not disturb, invisible, and offline. Kietzmann et al. (2011) note that “presence” can bridge the real and virtual worlds. Thus, SM sites such as

Foursquare provide locational information. Friends Around Me and similar SM sites are focused on geographical spaces and can be synced with Gowalla, Foursquare, Facebook, and Twitter, sites that allow for many of the operations mentioned above including profile development, conversations, and sharing of photos and virtual gifts. The Carbon Project software company has trademarked the term “geosocial networking.” User availability in time and space, as noted above, is vital for location-based services used for commercial and emergency notifications.

Relationships specify how users are linked to others. Those using LinkedIn to request a new connection are questioned as to whether they are a colleague, classmate, business contact, friend, or other. Clicking on “other” will require an email address of the intended contact to be provided before a request to connect is sent. Conversely, LinkedIn provides for the ability to see how others are connected and the degrees of separation between the user and his or her intended contacts. Relationships themselves may be characterized by the two attributes of structure and flow. These may be seen as two competing “camps.” Structure refers to the size, density, and centrality of an individual’s link within their social graph. This is the very heart of the social network science that has been developed over the decades and which was summarized in the work of Wasserman and Faust and Scott, discussed above. Flow has attracted more interest in the last decade and is concerned with how user relationships are defined by the use, exchange, and transformation of tangible and intangible resources between individuals in the SN.

Reputation – for relationships to be effective, processes that validate the authenticity and reputation of the users must be established. To assist in this process, various social metrics have been suggested. These metrics fall into two categories: empirical metrics that are based on, for example, simple measures such as the number of followers an individual has on Twitter or the number of “likes” a business has on Facebook and metrics based on mathematical formalizations (Nielsen and Krukow 2004). The most appropriate metric to measure reputation will vary depending on the individual, the business, or the SN website being used. Some social media sites, such as Social Mention, claim to track more than 100 SN sites so as to determine what is being said about a given individual, product, or business. Social Mention uses a number of metrics including strength, determined by the number of mentions; sentiment, measured by the ratio of positive to negative mentions; and reach, the numbers of different users that mention the target divided by the total number of mentions. Schubring (2012) assesses 12 different social media monitoring tools including Social Mention, some of which are free, while others such as Radian 6 may cost \$500/month or more depending on volume of traffic.

Groups – finally, Kietzmann et al. (2012) characterize SM by the extent to which users are ordered into categories or groups, and, as noted above, this is one reason that SM sites are “small worlds.” Groups may be user identified whereby individuals place their contacts into self-defined categories such as friends, business contacts, and interest groups in terms of hobbies or professional interests. Alternatively, groups may be similar to clubs in the offline world in the sense that they may be open to any member of the SN or might be by invitation

only or indeed might be secret. Many professional, regional science organizations including the North American Regional Science Council and the European Regional Science Association, for example, also have discussion groups in SN sites such as LinkedIn and can be “followed” on Facebook. In addition, sub-groups, such as NECTAR, within these organizations can also be joined. These groups usually have an open membership, but even so users must sign on to them. In this sense, they will be affected by the rules governing conversations mentioned above. Groups touch on almost all other aspects of SM sites, especially in the way they communicate, collaborate, and share and in the manner in which they develop trust and support among users.

38.6 The Development of Explicit Social Network Platforms

Since the founding of explicitly SN websites in the late 1990s (Boyd and Ellison 2007; Waters 2012), interest in social networking has developed with unsurpassed rapidity. Because of the speed with which the social networking world changes, the best sources of information are online resources such as Wikipedia. Indeed, since Wikipedia contributors are in some senses a social networked community, it would be ironic to ignore this resource. Wikipedia (2012) provides an alphabetized list of most of the existing SN sites including information on the emphasis of the site, the date when it was founded, how many registered users, whether registration is open or restricted and in what manner it is restricted, and how the sites are ranked by using a *page ranking* system.

Wikipedia also provides a list of *virtual communities* with over 100 million users. Although a virtual community is described as a “social network of individuals who interact through specific social media, potentially crossing geographical and political boundaries in order to pursue mutual interests or goals,” a number of these social networks are not included on Wikipedia’s previously mentioned list. These include Windows Live, Tencent Weibo, and Skype. Obviously, the definition of both a social network and a virtual community is somewhat fluid, and this is even more so when websites add to their services in an incremental fashion.

There are many ways of presenting and organizing information on social networking websites other than alphabetical lists. One of the more interesting approaches is simply to show the dominance of individual sites on a world map (Waters 2012) reflecting the influence of both language and national preferences. The map shows the dominating network in each country. The data was current as of February 2011. The map will continue to change rapidly. For instance by mid-2012, the number of Facebook users had risen from 640 million to more than 900 million. The map shows clearly that Facebook not only has the largest number of users but also the greatest global reach. While Qzone is the second most important SN site in numbers of users, it dominates only in China. Orkut, owned and operated by Google, is overwhelmingly dominant in Brazil. An aspatial group of SN websites by category is provided by the Social Media Influence website (Waters 2012).

38.7 Measuring Individual Influence Within Social Networks

Three companies, PeerIndex, Klout, and Kred that measure influence within a social network, use a combination of methods from social network analysis. Complete details of the algorithms are not revealed by the companies but are supposedly based on a combination of connections and activity. Perhaps the most detailed explanation of its methodology is provided by PeerIndex whose website states that on any given topic users' scores will reflect their authority (i.e., how much others rely on and trust their opinions and recommendations), their audience (size and responsiveness are important), and their activity (which is measured relative to the level of activity within their community and which should be consistent, i.e., neither frenetic nor spasmodic).

The PeerIndex website FAQ notes that “improving your scores is really pretty simple: share good and timely information, engage with authority figures in the topic, make sure your followers are largely real people, and we’ll take care of the rest.” Elsewhere, it is explained that topics only become viable for ranking when a group that is interested in a particular topic becomes both large and active.

The Kred website argues that they are the only site to measure influence within social networks that is fully transparent. Whether this is completely true or not is debatable. Kred measures both influence and outreach using both Twitter and Facebook activity, supposedly measuring trust and generosity. Influence is measured by a user’s ability to inspire action assessed on how frequently they are “Retweeted, Replied, Mentioned and Followed on Twitter.” Facebook interactions that count toward a user’s Kred include Facebook “Posts, Mentions, Likes, Shares and Event Invitations.” Outreach is assessed by a user’s “generosity in engaging with others” plus how often a user retweets, replies, or mentions others. Interactions on Facebook that “count” include “Posts, Mentions, Comments and Likes.” According to Klout, their scores reflect the true reach (how many people you influence), amplification (how much you influence them), and your network impact (the influence of your network).

Put simply, websites such as Alexa measure, in a general sense, the importance of a given social network, while sites such as PeerIndex, Klout, and Kred measure the influence of nodes, that is, individuals within those networks. This is akin to the work in the early days of network analysis where researchers measured the connectivity of both the entire network structure and the importance and connectivity of individual nodes within those networks.

38.8 Recent Developments in Network Science

A number of network science texts have been published in the last few years. For the regional scientist, two of the most important are Goyal (2007) and Hansen et al. (2011). The former contribution concentrates on economic applications of network science, the final chapters of the book providing a detailed treatment of labor markets, network formation, and research collaboration among firms; thus, spatial

influences are implied even if they are not discussed explicitly. The latter, edited text reviews the various subdisciplines and applications of network science, provides a series of case studies, and includes directions on how to use the NodeXL programming environment. SNs have many special characteristics such as grouping and clustering (noted above) that are not found to the same extent in technological and biological networks and as a consequence require specialized methods and techniques for revealing their structures. In addition to the SN handbooks discussed above, an Encyclopedia of Social Networks (Barnett 2011) has recently been published.

Specialized journals are being established on an ongoing basis, and important among these are Social Networks, Social Networks: An International Journal of Structural Analysis, International Journal of Virtual Computing and Social Networks, International Journal of Social Network Mining, Cyberpsychology, Behavior and Social Networking, Social Network Analysis and Mining, Network Science, Journal of Social Structure, The Journal of Mathematical Sociology, and Journal of Computer-Mediated Communication. Since 2008, SIGCOMM has organized an annual Workshop on Online Social Networks (WOSN) and since 2009 a Workshop on the Social Mobile Web (SMW).

38.9 The Decline of Distance and the Need for a Second Law of Geography

A primary attraction of joining a social network is that it supposedly weakens or even removes the constraints of distance. If this is true, then SNs must surely have a major impact on the spatial distribution of economic activity and will therefore be of great interest to regional scientists. To investigate these concerns, in December 2010 the Center for Spatial Studies at the University of California, Santa Barbara (UCSB), organized a workshop to determine a research agenda that would investigate the temporal and spatial constraints of SNs. Almost all the participants prepared position papers (and a final report was also issued for a review and web links, see Waters 2012). If there are no longer any spatial constraints, then Tobler's widely cited "*First Law of Geography*" (Tobler 2004) no longer applies and might be replaced by a *Second Law of Geography*: "Everything is connected to everything else, but things more closely connected are more related – and geography may well be irrelevant." Interestingly, when a debate was held on the First Law at an Association of American Geographers Conference and the discussion subsequently published in the annals of that organization, the "small world" literature was raised (Tobler 2004), but none of the commentators addressed the issue of the lack of spatial autocorrelation within an SN nor whether the pattern of SN memberships replicated real-world geography, distance decay, and spatial interaction. However, 6 years later this was a concern of the participants at the UCSB conference. It is to be expected that much of the ongoing research that will be conducted by regional scientists will be focused on this one issue.

Recent research into SNs suggests that for the most “social” of the SNs geography does matter. Barthelemy (2010) reviews a number of papers that document varying degrees of distance decay in real-world social networks. This might have been expected for mobile phone data where the probability that two individuals were connected was found to be proportional to the Euclidean distance between them raised to the minus 2 power (the classical version of the gravity model). Barthelemy’s discussion notes that in studies of the blogging SN, LiveJournal, on average users had eight friends of which 5.5 were geographically influenced and lived in close proximity with a distance decay function proportional to the inverse of the first power of distance, while the remaining 2.5 friends resulted from non-geographic processes. Another study reviewed by Barthelemy suggested that an exponent of approximately one is also appropriate for modeling the spatial separation of email correspondence of Live Blogger users and Facebook friends, respectively. In that study, the authors echoed Rietveld and Vickerman’s (2004) observation that “distance is not dead.” Further evidence in support of a strong geographical influence is provided by Scellato et al. (2010) for the BrightKite, Twitter, Foursquare, and LiveJournal SNs.

Thus, it may be concluded that both geography and other various and perhaps complex social processes will determine the links between friends in any given SN and that the purpose of a particular SN is likely to influence just how strong each component is. It might be expected that geography and distance would have a much weaker influence on connections in a business-oriented network such as LinkedIn. However, few comparisons of the geographies of Facebook and LinkedIn have appeared, and those that have been published have not explicitly addressed the spatial differences (Papacharissi 2009).

In a recent paper, Singleton and Longley (2009) have discussed the differences between online and offline geographic spaces, suggesting that much of the work on geodemographics that has been such a widespread and lucrative application of GIScience will now have to be recast so as to take into account the joint geographic and social aspects of SNs. The way forward would appear to be to adopt the new metrics developed by Scellato and his colleagues (Scellato et al. 2010) that include a “node locality” metric that measures the geographic closeness of the neighbors of a node and secondly a “geographic clustering coefficient” that measures how tightly connected the neighborhood of a node is based on the proportion of triangular links around a node where these are weighted by a distance decay function. It is measures such as these, which provide a link to earlier work on time-space geography that was originally developed by the Swedish geographer Torsten Hagerstrand that will allow regional scientists to determine the spatial, economic impact of SNs. The likelihood is that distance and geography will still have a strong role to play in any future regional science that makes use of social network analysis.

38.10 Conclusions

In the past, it became common wisdom that “The Internet Changes Everything.” Today, it can be argued with equal conviction that social networks will have

a similar impact on economic activity and that the three primary activities facilitated by the Internet, namely, access to and sharing of content (books, videos, and music), communication (email, instant messaging), and self-expression (blogs), can all be achieved through SNs. Indeed, it seems reasonable to suggest that Facebook itself does all of the above. That this will have a major impact on economic activity is now beyond dispute. The extent to which SNs will have a spatial impact and will attract the attention of regional scientists has yet to be determined and will be dependent on the degree to which they alter the spatial distribution of our activities. In the coming years, we can expect to see new books and extensive research on the spatial impacts of SNs.

References

- Barabasi A-L, Albert R (1999) Emergence of scaling in random networks. *Science* 286(5439):509–512
- Barnett GA (2011) Encyclopedia of social networks. Sage, Thousand Oaks
- Barthelemy M (2010) Spatial networks. *Phys Rep* 499(1–3):1–101
- Boyd DM, Ellison NB (2007) Social network sites: definition, history and scholarship. *J Comp Med Commun* 13(1):210–230
- Cairncross F (2001) *The death of distance: How the communications revolution is changing our lives* (second edition; first edition published 1997). Harvard Business School Press, Cambridge
- Christakis NA, Fowler JH (2011) Connected: the surprising power of our social networks and how they shape our lives – how your friends' friends' friends affect everything you feel, think and do. Back Bay Books, Little, Brown, New York
- Goyal S (2007) An introduction to the economics of networks. Princeton University Press, Princeton
- Granovetter M (1973) The strength of weak ties. *Am J Sociol* 78(6):1360–1380
- Haggett P, Chorley R (1969) Network analysis in geography. St. Martin's Press, New York
- Hansen DL, Shneiderman B, Smith MA (2011) Analyzing social media networks with NodeXL: insights from a connected world. Elsevier, Burlington
- Kaplan AM, Haenlein M (2010) Users of the world, unite! The challenges and opportunities of social media. *Bus Horiz* 53(1):59–68
- Kaplan AM, Haenlein M (2011) The early bird catches the news: nine things you should know about micro-blogging. *Bus Horiz* 54(2):105–113
- Kietzmann JH, Hermkens K, McCarthy IP, Silvestre BS (2011) Social media? Get serious! Understanding the functional building blocks of social media. *Bus Horiz* 54(3):241–251
- Markoff J, Sengupta S (2011) Separating you and me? 4.74 Degrees. Accessed on July 15th, 2012, at <http://www.nytimes.com/2011/11/22/technology/between-you-and-me-4-74-degrees.html>
- Nielsen M, Krukow K (2004) On the formal modelling of trust in reputation-based systems. In: Karhumaki J, Maurer H, Paun G, Rozenberg G (eds) *Theory is forever: essays dedicated to arto salomaa (salomaa festschrift)*, lecture notes in computer science, 3113th edn. Springer, Berlin, pp 192–204
- Papacharissi Z (2009) The virtual geographies of social networks: a comparative analysis of facebook, LinkedIn and ASmallWorld. *New Med Soc* 11(1–2):199
- Reed WJ (2001) The Pareto, zipf and other power laws. *Econ Lett* 74(1):15–19
- Rietveld P, Vickerman R (2004) Transport in regional science: the “death of distance” is premature. *Pap Reg Sci* 83(1):229–248
- Scellato S, Mascolo C, Musolesi M and Latora V (2010) Distance matters: geo-social metrics for online social networks. In Proceedings of the 3rd workshop of online social networks, WOSN 2010, Boston, USA (unpaginated). Berkeley, CA: USENIX Association

- Scott JP (1992) Social network analysis: a handbook. Sage, Newbury Park
- Scott JP (2011) The Sage handbook of social network analysis. Sage, Thousand Oaks
- Singleton AD, Longley PA (2009) Geodemographics, visualization, and social networks in applied geography. *Appl Geogr* 29(3):289–298
- Smith RHT, Taaffe EJ, King LJ (1969) Readings in economic geography. Rand McNally, New York
- Strogatz S (2005) Romanesque networks. *Nature* 433(7024):365–366
- Tobler WR (2004) On the first law of geography: a reply. *Ann Assoc Am Geogr* 94(2):304–310
- Travers J, Milgram S (1969) An experimental study of the small world problem. *Sociometry* 32(4):425–443
- Wasserman S, Faust K (1994) Social network analysis: methods and applications. Cambridge University Press, Cambridge
- Waters NM (2006) Network and nodal indices: measures of complexity and redundancy: a review. In: Reggiani A, Peter Nijkamp P (eds) Spatial dynamics, network and modelling. Edward Elgar, Cheltenham/Northampton/USA
- Waters NM (2012) Social networks: is spatial special when it's social? *GeoWorld*, 25, in press
- Watts DJ, Strogatz SH (1998) Collective dynamics of “small-world” networks. *Nature* 393(6710):440–442
- Watts DJ, Dodds PS, Newman MEJ (2002) Identity and search in social networks. *Science* 296(5571):1302–1305
- Willinger W, Alderson D, Doyle JC (2009) Mathematics and the internet: a source of enormous confusion and great potential. *Notices of the AMS* 56(5):586–599
- Wikipedia (2012) Social network analysis software. Accessed on July 12th, 2012, at http://en.wikipedia.org/wiki/Social_network_analysis_software

Michael Wegener

Contents

39.1	Introduction	742
39.2	Theory	743
39.3	Operational Models	746
39.3.1	Spatial-Interaction Location Models	747
39.3.2	Accessibility-Based Location Models	748
39.4	Current Debates	750
39.4.1	Equilibrium or Dynamics	750
39.4.2	Macro or Micro	752
39.5	Future Challenges	753
39.6	Conclusions	755
	References	756

Abstract

The relationship between urban development and transport is not simple and one way but complex and two way and is closely linked to other urban processes, such as macroeconomic development, interregional migration, demography, household formation, and technological innovation. In this chapter, one segment of this complex relationship is discussed: the two-way interaction between urban land use and transport within urban regions. The chapter looks at integrated models of urban land use and transport, i.e., models that explicitly model the two-way interaction between land use and transport to forecast the likely impacts of land use and transport policies for decision support in urban planning. The discussion starts with a review of the main theories of land-use transport interaction from transport planning, urban economics, and social geography.

M. Wegener

Spiekermann & Wegener, Urban and Regional Research, Dortmund, Germany

e-mail: mw@spiekermann-wegener.de

It then gives a brief overview of selected current operational urban models, thereby distinguishing between spatial-interaction location models and accessibility-based location models, and discusses their advantages and problems. Next, it reports on two important current debates about model design: are equilibrium models or dynamic models preferable, and what is the most appropriate level of spatial resolution and substantive disaggregation? This chapter closes with a reflection of new challenges for integrated urban models likely to come up in the future.

39.1 Introduction

The history of urban settlements is closely linked to transport. Cities appeared in human history when technological innovation required the spatial division of labor between specialized crafts and agricultural labor and gave rise to urban–rural travel and goods transport. Cities were established at trade routes, ports, or river crossings and became origins and destinations of trade flows. Cities were compact, as all movements were done on foot, until the railway and later the automobile opened the way to today's sprawling agglomerations.

These brief notes already show that the relationship between urban development and transport is not simple and one way but complex and two way. On the one hand, spatial division of labor, i.e., the separation of locations of human activities in space, requires spatial interaction, i.e., travel and goods transport. On the other hand, the availability of transport infrastructure, such as roads, railways, and airlines, makes locations attractive as residences or business locations and so affects real estate markets and the choice of location of households and firms. Moreover, it becomes clear that the relationship between urban development and transport is closely linked to other urban processes, such as macroeconomic development, interregional migration, demography and household formation, and technological innovation.

In this chapter, one segment of the complex relationship between urban development and transport is discussed: the two-way interaction between urban land use and transport within urban regions. The macroeconomic dimension dealing with growth or decline of whole cities within urban systems is addressed in several other chapters, such as ► Chaps. 45, “[Interregional Input–Output Models](#),” ► 46, “[Interregional Trade Models](#)”.

This chapter looks at *integrated* models of urban land use and transport, i.e., models which explicitly model the two-way interaction between land use and transport to forecast the likely impacts of land-use policies, such as zoning or building density or height constraints, and of transport policies, such as transport infrastructure investments, public transport improvements, or taxes or user charges, for decision support in urban planning. That excludes transport models per se which predict traffic patterns that result from different land-use configurations and land-use change models that predict likely land-use changes that result from a particular transport system, as well as models that deal only with one urban subsystem, such as housing or business location.

The discussion proceeds from a review of the main theoretical approaches of land-use transport models and a brief overview of operational models to current debates and new challenges that are likely to influence future development in this field.

There are in the literature several reviews of integrated land-use transport models, such as Wegener (2004) and Hunt et al. (2005).

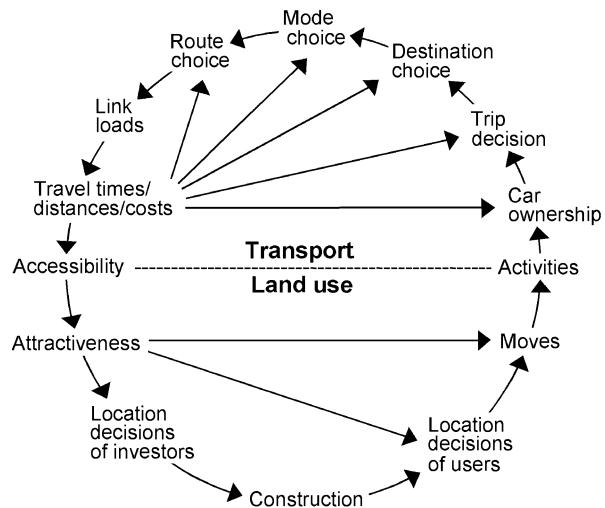
39.2 Theory

Urban land-use transport models originated in the United States in the 1960s as part of the diffusion of operations research and systems theory into all fields of society. The first attempts to model the interaction between land use and transport were initiated by transport planners who felt that predicting future traffic flows without taking account of their impacts on location was inadequate. Hansen (1959) showed for Washington, DC, that locations with good accessibility had a higher chance of being developed, and at a higher density, than remote locations (“how accessibility shapes land use”). The recognition that mobility and location decisions co-determine each other and that therefore transport and land-use planning need to be coordinated led to the notion of the “land-use transport feedback cycle”. The set of relationships implied by this term can be summarized as follows (Wegener and Fürst 1999, see Fig. 39.1):

- The distribution of *land uses*, such as residential, industrial, or commercial, over the urban area determines the locations of households and firms and so the locations of human *activities* such as living, working, shopping, education, and leisure.
- The distribution of human *activities* in space requires spatial interactions or trips in the *transport system* to overcome the distance between the locations of activities.
- These spatial interactions are based on decisions of travelers about car availability, number of trips, destination, mode, and route. They result in *traffic flows* and, in case of congestion, in increased travel times, trip lengths, and travel costs.
- Travel times, trip lengths, and travel costs create opportunities for spatial interactions that can be measured as *accessibility*.
- The spatial distribution of accessibility influences, among other attractiveness indicators, location decisions of investors and results in changes of the *building stock* by demolition, upgrading, or new construction.
- These changes in building supply determine location and relocation decisions of households and firms and thus the distribution of *activities* in space.

This simple explanation pattern is used in many engineering-based and human-geography urban development theories. These start from origins and destinations, such as workers and workplaces, and from these infer trip volumes that best reproduce observed trip frequency distributions. It had already been observed by Ravenstein (1885) and Zipf (1949) that the frequency of human interactions, such as messages, trips, or migrations between two locations (cities or regions), is proportional to their size but inversely proportional to their distance. The analogy to the law of gravitation in physics is obvious.

Fig. 39.1 The land-use transport feedback cycle (Wegener and Fürst 1999, 6)



The gravity model was the first *spatial-interaction* model. Its physical analogy has later been replaced by better founded formulations derived from statistical mechanics (Wilson 1967) or information theory (Snicksars and Weibull 1977). Only later did it become possible (Anas 1983) to link it via random utility theory (Domencich and McFadden 1975) to psychological models of human decision behavior.

From the spatial-interaction model, it is only a small step to its application as a location model. If it is possible to draw conclusions from the spatial distribution of human activities to the interactions between them, it must also be possible to identify the location of activities giving rise to a certain trip pattern. Wilson (1970) distinguishes four types of urban *spatial-interaction location* models: unconstrained models, production-constrained models, attraction-constrained models, and doubly constrained models. Unconstrained models deal with households without fixed residence or workplace, production-constrained models with households looking for a job, and attraction-constrained models with households looking for a residence. The doubly constrained model is actually not a location model but the familiar transport model (see ▶ Chap. 36, “Travel Behavior and Travel Demand”).

To give an example, the production-constrained spatial-interaction model is written as follows:

$$T_{ij} = A_i O_i D_j \exp(-\beta c_{ij}) \quad (39.1)$$

$$A_i = 1 / \sum_j D_j \exp(-\beta c_{ij}) \quad (39.2)$$

$$p_{ij} = \frac{D_j \exp(-\beta c_{ij})}{\sum_j D_j \exp(-\beta c_{ij})} \quad (39.3)$$

where T_{ij} are trips between zone i and zone j , O_i are trips generated by i and D_j trips attracted by j , and c_{ij} is the travel time or travel cost, or both, between i and j . The β is a parameter indicating the sensitivity to travel cost; because of its negative sign, more distant destinations are less likely to be selected. A_i is the so-called balancing factor ensuring that total trips equal O_i , and p_{ij} is the probability that a trip goes from i to j .

A second set of theories focuses on the *economic* foundations of land use. A fundamental assumption of all spatial economic theories is that locations with good accessibility are more attractive and have a higher market value than peripheral locations. This assumption goes back to von Thünen (1826) and has since been varied and refined in many ways (see ► Chap. 27, “[Classical Contributions: Von Thünen, Weber, Christaller, Lösch](#)”). Probably the most influential example of the latter kind is the model of the urban land market by Alonso (1964). The basic assumption of the Alonso model is that firms and households choose that location at which their bid rent, i.e., the land price they are willing to pay, equals the asking rent of the landlord, so that the land market is in equilibrium. The bid rent of firms results from the cost structure of their production function, i.e., sales price minus production and transport costs plus profit divided by size of land. A firm having a higher added value per unit of land is therefore able to pay a higher price than a firm with less intensive land utilization, everything else being equal. So it is not surprising that, say, jewelers are found in the center, whereas trucking companies have their yards on the periphery. Alonso’s model has been the point of departure for a multitude of urban-economics model approaches. In more advanced variations of the model, restrictive assumptions, such as the monocentric city or perfect competition and complete information, have been relaxed (e.g., Anas 1982).

A third group of theories used in land-use transport models are *social* theories. In social theories of urban development, the spatial development of cities is the result of individual or collective appropriation of space. Based on an adaptation of evolutionist thoughts from philosophy (Spencer) and biology (Darwin), the Chicago school of urban sociologists interpreted the city as a multispecies ecosystem, in which social and economic groups fight for ecological positions. Appropriation of space takes place in the form of immigration of different ethnic or income groups or tertiary activities into residential neighborhoods, and concepts of animal and plant ecology, such as “invasion,” “succession,” or “dominance,” are used to describe the phases of such displacement.

Social geography theories go beyond the macro perspective of social ecology by referring to age-, gender-, or social-group specific activity patterns which lead to characteristic spatiotemporal behavior and hence to permanent localizations. Action space analyses (e.g., Chapin and Weiss 1968) identify the frequency of performance of activities reconstructed from daily space-time protocols as a function of distance to other activities and draw conclusions from this for the most probable allocation of housing, workplaces, shopping, and recreation facilities or, in other words, for the most likely level of spatial division of labor in cities.

Hägerstrand (1970) made these ideas operational by the introduction of “time budgets,” in which individuals, according to their social role, income, and level of

technology (e.g., car ownership), command action spaces of different size and duration subject to three types of constraints: (i) *capacity constraints*, i.e., personal, nonspatial restrictions on mobility, such as monetary budgets, time budgets, availability of transport modes, and ability to use them; (ii) *coupling constraints*, i.e., restrictions on the coupling of activities by location and time schedules of facilities and other individuals; and (iii) *institutional constraints*, i.e., restrictions of access to facilities by public or private regulations such as property, opening hours, entrance fees, or prices. Only locations within the action spaces can be considered as destinations or permanent locations.

On the basis of Hägerstrand's action space theory, Zahavi (1974) proposed the hypothesis that individuals in their daily mobility decisions do not, as the conventional theory of travel behavior assumes, *minimize* travel time or travel cost needed to perform a given set of activities but instead *maximize* activities or opportunities that can be reached within their travel time and money budgets.

39.3 Operational Models

Lowry's (1964) *Model of Metropolis* was the first attempt to quantify the land-use transport feedback cycle in one integrated model. The model consists of two singly constrained spatial-interaction location models, a residential location model and a service and retail employment location model, nested into each other. In modern notation, the two models would be written as

$$T_{ij} = \frac{R_i \exp(-\beta c_{ij})}{\sum_i R_i \exp(-\beta c_{ij})} E_j \quad (39.4)$$

$$S_{ij} = \frac{W_j \exp(-\beta c_{ij})}{\sum_i W_j \exp(-\beta c_{ij})} P_i \quad (39.5)$$

where T_{ij} are work trips between residential zone i and work zone j and S_{ij} shopping trips between residential zone i to retail facilities in zone j . E_j are workers in j and P_i population in i to be distributed, and R_i are dwellings in i and W_j shopping facilities in j used as destinations in the two spatial-interaction models, and c_{ij} is the travel time between i and j . In the first iteration, only work trips to the workplaces of basic industries, i.e., industries exporting to other regions and not serving the local population, are modeled. The two spatial-interaction location models are linked by assumptions about how many people are supported by one worker and how many retail employees are supported by one resident. In each subsequent iteration, workers and residents are updated until they no longer change, i.e., until the system is in equilibrium.

The Lowry model stimulated a large number of increasingly complex land-use transport models in the USA and not much later also in Europe. Many of these early models were not successful because of unexpected difficulties of data collection

and calibration and the still imperfect computer technology of the time. More important, however, was that the models were mainly oriented toward urban growth and the efficiency of the transport system and had nothing to say about the ethnic and social conflicts arising in US cities at that time. Moreover, the models were committed to the paradigm of synoptic rationalism in planning theory, which was increasingly replaced by incremental, participatory forms of planning. In his “Requiem for Large Scale Models,” Lee (1973) accused the models of “seven sins”: hypercomprehensiveness, grossness, mechanicalness, expensiveness, hungeriness, wrongheadedness, and complicatedness.

But many of the technical problems of the early models were solved by better data availability and faster computers. The spatial and substantial resolution of the models was increased, and they were based on better theories, such as bid-rent theory, discrete choice theory, and user equilibrium in transport networks (see ► Chap. 40, “Network Equilibrium Models for Urban Transport”). In addition, better visualization techniques made the results of the models better understood by citizens and policy makers. A new generation of models paid more attention to aspects of social equity.

The 1990s brought a revival in the development of urban land-use transport models. New environmental legislation in the USA required that cities applying for federal funds for transport investments demonstrate the likely impacts of their projects on land use. This had the effect that virtually all major metropolitan areas in the USA maintained an integrated land-use transport model. In Europe, the European Commission initiated a large research program *The City of Tomorrow*, in which integrated land-use transport models were applied in several research projects (Marshall and Banister 2007). Several integrated land-use transport models were applied in a growing number of metropolitan areas. New developments in data availability brought about by geographical information systems (GIS) and further advances in computer technology have removed former technical barriers.

It is impossible to present here all operational integrated land-use transport models existing in the world today. Instead a classification of models by the way they implement the feedback from transport to land use is proposed using a few examples, recognizing that in each group, there exists a great variety of approaches.

39.3.1 Spatial-Interaction Location Models

Spatial-interaction location models retain the original Lowry concept by modeling the location of human activities as destinations of trips using the production-constrained spatial-interaction model. The most prominent urban model of this kind still operational today is the MEPLAN model developed by Echenique (1985) as well as its offsprings, TRANUS (de la Barra 1989) and PECAS (Hunt and Abraham 2005). All three models use a multi-industry, multiregional input–output framework (see ► Chap. 45, “Interregional Input–Output Models”) to predict the locations of production and consumption in the urban region, where households of different types are treated as industries producing labor and

consuming commodities. By iterating between the land-use parts and the transport parts of the models, general equilibrium between transport costs (including congestion) and land and commodity prices is achieved. The core equation of MEPLAN is

$$X_{irs} = X_{ir} A_{ir} f(c_{ir} + g_{irs}) Z_{is} \quad (39.6)$$

where X_{irs} are deliveries of industry i from region r to region s , X_{ir} is the supply of goods of industry i in r and Z_{is} the demand for such products in s , and c_{ir} are unit production costs of such products in r and g_{irs} their unit transport costs from r to s . A_{ir} is the balancing factor as in Eq. (39.1) ensuring that total trade flows from region r equal production in r .

The great advantage of spatial-interaction location models is their firm foundation in economic theory with respect to production and consumption. One possible criticism is that households are treated as industries producing labor and consuming commodities, with the consequence that residential location solely depends on workplace location, as if workers decided where to live on their way back from work.

In his most recent model RELU-TRAN, Anas reverses the causal direction of the input–output framework by modeling the location choice of consumers (households), producers (firms), landlords, and developers separately by utility-based production functions which include for households the costs of budget-constrained trips and for firms interindustry links as generated by the transport part of the model. As in the input–output models, by iterating between the land-use and transport parts of the model, general equilibrium between land use and transport is achieved (Anas and Liu 2007).

39.3.2 Accessibility-Based Location Models

The second group of land-use transport models predicts not actual spatial interactions but the opportunity for spatial interactions at potential locations. The indicator of opportunity for spatial interactions is called accessibility. Accessibility indicators can take a wide range of forms, from simple accessibility indicators, such as distance to the nearest bus station or motorway exit, to complex indicators measuring the ease of reaching all destinations of interest. The most frequently used complex accessibility indicator is potential accessibility or the total of all destinations of interest weighted by an inverse function of the effort to reach them measured in time or cost or a combination of both as “generalized cost”:

$$A_i = \sum_j D_j \exp(-\beta c_{ij}) \quad (39.7)$$

where A_i is the potential accessibility of zone i with respect to destinations of interest D_j and c_{ij} is the generalized costs of travel between i and j . The inverse similarity with the balancing factor of Eq. (39.2) is obvious.

Examples of operational accessibility-based location models in use today are IRPUD (Wegener 1982), RURBAN (Miyamoto and Udomsri 1996), MUSSA (Martinez 1996), DELTA (Simmonds 1999), and UrbanSim (Waddell 2002). These models predict location choices of households and firms with discrete choice models using multi-attribute utility functions in which accessibility indicators are combined with other attributes of potential locations to indicate their attractiveness from the point of view of households looking for a residential location or firms looking for a business location. In that respect, these models build on the bid-rent approach of Alonso (1964), although equilibrium between asking rents and bid rents on the land market is achieved only in MUSSA, whereas the other three models keep land prices fixed during a simulation period and defer the price response of landlords to the next simulation period.

As an example of accessibility-based location choice, the allocation of housing demand to vacant residential land by a multinomial logit model in the IRPUD model is shown (Wegener 2011a):

$$C_{kli}(t, t+1) = \frac{L_{kli} \exp[\beta_k u_{kli}(t)]}{\sum_{il} L_{kli} \exp[\beta_k u_{kli}(t)]} C_k(t, t+1) \quad (39.8)$$

where $C_k(t, t+1)$ are new dwellings of type k developers plan to build in the whole region between time t and $t+1$, $C_{kli}(t, t+1)$ are dwellings of that type that will be built on land-use category l in zone i in that period, and L_{kli} is the capacity of vacant land for such dwellings given zoning and building density and height constraints. The parameters β_k indicate the selectivity of developers with respect to the attractiveness $u_{kli}(t)$ of land-use category l in zone i for dwellings of housing type k :

$$u_{kli}(t) = [u_{ki}(t)]^{v_k} [u_{kl}(t)]^{w_k} [u(c_{kli})(t)]^{1-v_k-w_k} \quad (39.9)$$

where $u_{ki}(t)$ is the attractiveness of zone i as a location for housing type k , $u_{kl}(t)$ is the attractiveness of land-use category l for housing type k , and $u(c_{kli})(t)$ is the attractiveness of the land price of land use category l in zone i in relation to the expected rent or price of the dwelling. The v_k , w_k , and $1 - v_k - w_k$ are multiplicative weights adding up to unity. The zonal attractiveness $u_{ki}(t)$ is multi-attribute and contains, besides other indicators of neighborhood quality, one or more types of accessibility indicators.

The advantage of accessibility-based location models is that by inserting different types of accessibility indicators into the utility functions of different types of locators, the great diversity of accessibility needs reflecting different lifestyles and preferences of households and different communication and transport needs of firms can be considered. Their disadvantage is that the actual travel and transport behavior, and hence actual travel times and transport cost, become known only in the next iteration of the associated transport model, but this may be acceptable because they change over time only gradually. The separation of the land-use and

transport parts of the model by the accessibility interface makes it easier to develop custom-tailored submodels of the location behavior of individual groups of actors, such as households looking for a dwelling, landlords looking for a tenant, developers considering upgrading of their housing stock or looking for vacant land for new residential buildings, or firms looking for vacant floorspace or for land to build new floorspace.

This has important implications for the software organization of the models. While spatial-interaction location models as described in the previous section tend to be “unified,” i.e., to consist of one single complex algorithm designed to achieve general equilibrium, the accessibility-based models described in this section tend to be “composite,” i.e., to consist of several interlinked modules each serving a specific purpose, modeling the behavior of a particular group of actors and using the accessibility indicators most appropriate for that.

39.4 Current Debates

The urban models sketched so far represent the main model types coexisting until the end of the 1990s. However, from then on, the urban modeling scene has become increasingly fragmented along two dividing lines. The first divide runs between equilibrium modeling approaches and models that attempt to capture the dynamics of urban processes. The second more recent divide runs between aggregate macro-analytic approaches and new microscopic agent-based models.

39.4.1 Equilibrium or Dynamics

The first urban models were static equilibrium models, such as the Lowry model which generated an “instant metropolis” at a point in time in the future. This tradition was maintained and is still strong in urban-economics models based on the notion that all markets, including urban housing, real estate, and transport markets, tend to move toward equilibrium between demand and supply and that therefore the equilibrium state is the most appropriate guidance for urban planning.

In contrast to this view, a different movement in urban modeling has become more interested in the adjustment processes going on in cities that may lead to equilibrium but more frequently do not. The proponents of this movement, influenced by systems theory and complexity theory, argue that cities have evolved over a long time and display a strong inertia which resists sudden changes toward a desired optimum or equilibrium (see ► Chap. 69, “Spatial Dynamics and Space-Time Data Analysis”). Following this view, urban change processes can be classified as slow, medium speed, and fast (Wegener et al. 1986):

- Slow Processes: *Construction*. Urban transport, communications, and utility networks are the most permanent elements of the physical structure of cities. The *land-use* distribution is equally stable; it changes only incrementally.

Buildings have a life-span of up to 100 years and take several years from planning to completion.

- Medium-Speed Processes: *Economic, Demographic, and Technological Change*. The most significant kind of *economic* change are changes in the number and sectoral composition of employment. *Demographic* changes affect population through births, ageing, and death and households through household formation and dissolution. *Technological* change affects all aspects of urban life, in particular transport and communication. These changes do not affect the physical structure of the city but the way it is used.
- Fast Processes: *Mobility*. There are even more rapid processes that are planned and completed in less than a year's time. They refer to the mobility of people, goods, and information within and between given buildings and communication facilities. These changes range from job relocations and residential moves to the daily pattern of trips and messages.

The advocates of dynamic models argue that in order to make realistic forecasts, it is necessary to explicitly take account of the different speeds of processes. In particular, they criticize the implicit assumption of spatial-interaction location models that households and firms are perfectly elastic in their location behavior and change to the equilibrium spatial configuration as if there were no transaction costs of moving.

In contrast, dynamic urban models make the evolution of the urban system through time explicit. Early dynamic urban models (Harris and Wilson 1978; Allen et al. 1981) treated time as a continuum. Today the most common form are recursive or quasi-dynamic models in which the end state of one simulation period serves as the initial state of the subsequent period. The length of the simulation period, usually 1 year, is the implicit time lag of the model, as changes occurring in one simulation period affect other changes only in the next simulation period. By using results from earlier simulation periods, the modeler can implement longer delays and feedbacks. For instance, if it is assumed that it typically takes 3 years to plan and build a house, a delay of 3 years between residential investment decisions and the new dwellings appearing on the market would be appropriate. Similar delays between investment decision and completion allow to model the typical cycles of over- and undersupply of office space.

Most current dynamic urban models are composite models, i.e., operate with a combination of custom-tailored submodels for different urban change processes. By selecting the sequence in which these submodels are processed during a simulation period, the modeler can give certain processes priority access to scarce resources. It is no coincidence that most dynamic land-use models are accessibility-based location models, i.e., use accessibility indicators as link between transport and land use and so take advantage of the possibility to select different types of accessibility for different types of development.

Most existing equilibrium urban models, however, are unified, i.e., apply one algorithm to all its parts, such as spatial-interaction location in the case of MEPLAN, TRANUS, and PECAS, or bid-rent location in the case of MUSSA,

because they aim at general equilibrium between supply and demand, which is easier to achieve in a unified model. However, the growing success of dynamic or quasi-dynamic models has had its effects on equilibrium models. Some spatial-interaction location models, such as MEPLAN and PECAS, have been made recursive, i.e., they are processed not only for a distant target year but for years in between and have been complemented by developer submodels producing residential, commercial, and industrial floorspace that serve as constraints for the allocation of households and economic activity in the equilibration of the subsequent simulation period.

39.4.2 Macro or Micro

The second major divide appearing in the urban modeling scene concerns the debate about the most appropriate level of spatial and substantive disaggregation.

The first urban models were zone-based like the travel models of the time, as the data required by both types of models were available only for relatively large statistical areas. However, in the 1990s, the growth in computing power and the availability of GIS-based disaggregate data fuelled by non-modeling applications, such as data capture, mapping, spatial analysis, and visualization, has had its impact on urban modeling. New modeling techniques, such as cellular automata (CA) and agent-based models developed and applied in the environmental sciences, were proposed for modeling land-use changes of high-resolution grid cells (see ► [Chap. 62, “Cellular Automata and Agent-Based Models”](#)). In transport planning, activity-based models modeling no longer trips but activity-related multi-stop tours have become the state of the art (see ► [Chap. 37, “Activity-Based Analysis”](#)). The impact of these developments on urban modeling has been a massive and still continuing trend toward disaggregation to the individual level or microsimulation.

There are important conceptual reasons for microsimulation, such as improved theories and growing knowledge about human cognition, preferences, behavior under uncertainty and constraints, and interactions between individuals in households, groups, and social networks (see ► [Chap. 38, “Social Network Analysis”](#)), a growing potential for individualization; the choice of diversified lifestyles and hence mobility and location patterns. Disaggregate models of individual behavior are better suited to capture this heterogeneity.

Microsimulation was first used in the social sciences by Orcutt et al. (1961). Early applications with a spatial dimension covered a wide range of processes, such as spatial diffusion and urban expansion (see ► [Chap. 63, “Spatial Microsimulation”](#)). Since the 1980s, several microsimulation models of urban land use and transport have been developed, such as the pioneering ILUTE (Salvini and Miller 2005). Stimulated by the technical and conceptual advances discussed above, agent-based microsimulation urban models are proliferating all over the world, including microsimulation versions of originally aggregate models, such as IRPUD, DELTA, and UrbanSim.

However, not all disaggregate urban modeling projects have been successful (see, for instance, Wagner and Wegener 2007; Nguyen-Luong 2008). Many large modeling projects had to reduce their too ambitious targets. The reasons for these failures are partly practical, such as large data requirements and long computing times, but partly also conceptual.

The most important conceptual problem is the lack of stability of microsimulation models due to stochastic variation. Stochastic variation, also called microsimulation or Monte Carlo error, is the variation in model results between simulation runs with different random number seeds (see ► Chap. 63, “[Spatial Microsimulation](#)”). In agent-based models of choice behavior, the magnitude of stochastic variation is a function of the ratio between the number of choices and the number of alternatives and the selectivity of the choosing agents (the β parameter in the equations of this chapter). The stochastic variation is small when a large number of agents with clear preferences choose between few alternatives, e.g., travel modes. It is large when a small number of agents with less pronounced preferences choose between a large number of alternatives, e.g., locations, such as grid cells, parcels, or zones, as in the case of residential or business location. In that case, the stochastic noise may be larger than the differences between competing planning alternatives under investigation, and the results may convey an illusionary sense of precision (Wegener 2011b).

There are several ways to overcome this dilemma, such as averaging the results to a higher spatial level or to artificially increasing the number of choices in the model. The most frequently recommended method is to run the model several times and to average across the results of the different runs, something rarely done because of the already long computation times of microsimulation models.

In conclusion, the microsimulation community has yet to find a proper answer to the stochastic variation problem. The optimum level of disaggregation may not be the most disaggregate one. What is needed is a theory of multilevel urban models to identify the appropriate level of conceptual, spatial, and temporal resolution for each modeling task.

39.5 Future Challenges

The world is changing fast, and so are the problems of urban planning. The first land-use transport models were growth-oriented and mainly addressed technical problems, such as the reduction of urban sprawl and traffic congestion. The second generation of models increasingly considered equity aspects, such as social and ethnic segregation, accessibility of public facilities, and distributive issues, such as who gains and who loses if certain policies are implemented. Today the third generation of models tries to take account of the observed individualization of lifestyles and preferences by ever greater spatial, temporal, and substantial disaggregation.

However, today new challenges are becoming visible that cannot be handled by many of the urban land-use transport models existing today.

The first challenge is to extend the models from land-use transport interaction models to land-use transport environment models. Today only few urban models are linked to environmental models to show the impacts of planning policies on greenhouse gas emissions, air quality, traffic noise, and open space (Lautso et al. 2004). As environmental submodels predicting air quality or noise propagation require high-resolution grid cell data, this model extension may give a new twist to the macro versus micro debate toward multilevel models using different spatial levels with different resolutions and upward and downward feedbacks. Even fewer models are able to model the reverse relationship, the impact of environmental quality, such as air quality or traffic noise, on location.

The second challenge is the transition from population growth to population decline already observed and foreseeable in many European cities. With small population decline and moderate economic growth, there is still demand for new housing because of decreasing household size and increasing floorspace per capita. The same is true for work places due to growing floorspace demand per worker. However, if the losses of population and employment become larger than the growth in floorspace demand per capita or per worker, the task is no longer the allocation of growth but the management of decline by new types of policies, such as rehabilitation of neighborhoods, upgrading of rundown housing, or conversion or demolition of derelict or vacant buildings. Only few current urban models are able to handle this.

The third and greatest challenge arises from the possibility of future energy crises and the requirements of climate protection. Both causes are likely to make mobility significantly more expensive. For model design, it does not matter whether car trips become more expensive through higher prices of fossil fuels on the world market or through government policies to meet greenhouse gas reduction targets. What matters is that these targets cannot be achieved without rigorous changes in the framework conditions of land use and transport in urban areas, in particular without significant increases in the price of fossil fuels.

Most current urban models are not prepared for this. Many of them are not able to model transport policies, such as carbon taxes, emissions trading, road pricing, or alternative vehicles and fuels, or land-use policies, such as strict development controls, improvement of the energy efficiency of buildings, or decentralized energy generation. Even fewer models are able to identify population groups or neighborhoods most affected by such policies or possible problems with access to basic services, such as schools or health facilities, or participation in social and cultural life in low-density suburban or rural areas.

Many current transport models cannot correctly predict the impacts of substantial fuel price increases. Many do not consider travel costs in modeling car ownership, trip generation, trip distribution, and modal choice. Many do not forecast induced or suppressed trips. Many use price elasticities estimated in times of cheap energy. Many do not consider household budgets for housing and travel.

Action space theory with explicit travel time and travel cost budgets permits to predict what will happen if speed and cost of travel are changed by environment-oriented planning policies. Acceleration and cost reduction in transport lead to

more, faster, and longer trips; speed limits and higher costs to fewer, slower, and shorter trips. In the long run, this has effects on the spatial structure. Longer trips make more dispersed locations and a higher degree of spatial division of labor possible; shorter trips require a better spatial coordination of locations. However, making travel slower and more expensive does not necessarily lead to a reconcentration of land uses back to the historical city center. In many urban regions, population has already decentralized so much that further deconcentration of employment would be more effective in achieving shorter trips than reconcentration of population.

That plausible forecasts of the impacts of substantial energy price increases can be made with land-use transport models based on action space theory was demonstrated by the results of the EU project *Scenarios for the Transport System and Energy Supply and their Potential Effects* (STEPs). They show that with appropriate combinations of transport and land-use policies, significant reductions in greenhouse gas emissions can be achieved without unacceptable loss of quality of life (Fiorello et al. 2006).

39.6 Conclusions

After half a century of development, there exists today a broad spectrum of mathematical models to predict the spatial evolution of cities subject to exogenous trends and land-use and transport policies. These models build on a range of theories from transport planning, urban economics, and social geography to explain the complex two-way interaction between urban land use and transport, i.e., the location of households and firms and the resulting mobility patterns in urban regions subject to concurrent economic, demographic, and technological developments. Stimulated by advances in data availability, theory development and computing technology, these models have reached an impressive level of sophistication and operational applicability.

However, the urban modeling field has recently become divided into camps with different modeling philosophies. In particular, two dividing lines are becoming visible: One is the divide between equilibrium approaches which assume that cities are essentially markets moving toward equilibrium between demand and supply and dynamic approaches focusing on adjustment processes of different speeds. The other is the divide between macro approaches dealing with statistical aggregates at the level of zones and micro approaches modeling individual households and firms at the level of grid cells or parcels. In each of the two debates, the advantages and disadvantages of the competing approaches are obvious, but what is missing is an open and honest assessment of their relevance for the validity and robustness of the results of the models. Collaborative research projects in which different models are applied to identical problems and their results compared by meta-analyses are still the exception.

A second issue regarding the future of urban models is the new challenges for urban planning. The growing importance of environmental impacts of land-use and transport policies has not yet fully been embraced by most urban models.

Neither has the transition from population growth to population decline already observed or foreseeable in many cities, a great challenge for some models originally designed for allocating growth. But the greatest challenge for urban models will be how to cope with the combined effects of future energy scarcity and the imperatives of climate change. During and after the energy transition, energy for transport and building heating will no longer be abundant and cheap but scarce and expensive. This will have fundamental consequences for mobility and location. Land-use transport models which are calibrated on behavior observed in times of cheap energy and do not consider the costs of travel and location in relation to household income cannot adequately forecast these consequences. To deal with significantly rising energy costs, land-use transport models must consider the basic needs of households which can be assumed to remain relatively constant over time, such as shelter and security at home, accessibility of work, education, retail and necessary services, and the constraints on housing and travel expenditures by disposable household incomes.

To avoid the danger that the models, as in the 1970s, are again rejected by the planning practice, they must give up some long-standing traditions and be prepared to adopt new modeling principles: less extrapolation of past trends but more openness to fundamental change, less reliance on observed behavior but more theory on needs, less consideration of preferences and choices but more taking account of constraints, and less effort on detail but more focus on basic essentials.

References

- Allen PM, Sanglier M, Boon F (1981) Models of urban settlement and structure as self-organizing systems. US Department of Transportation, Washington, DC
- Alonso W (1964) Location and land use. Harvard University Press, Cambridge, MA
- Anas A (1982) Residential location models and urban transportation: economic theory, econometrics, and policy analysis with discrete choice models. Academic, New York
- Anas A (1983) Discrete choice theory, information theory and the multinomial logit and gravity models. *Transportation Res B* 17(1):13–23
- Anas A, Liu Y (2007) A regional economy, land use and transportation model (RELU-TRAN): formulation, algorithm design and testing. *J Regional Sci* 47(3):415–455
- Chapin FS, Weiss SF (1968) A probabilistic model for residential growth. *Transportation Res* 2(4):375–390
- de la Barra T (1989) Integrated land use and transport modelling. Cambridge University Press, Cambridge
- Domencich TA, McFadden D (1975) Urban travel demand: a behavioral analysis. North Holland, Amsterdam
- Echenique MH (1985) The use of integrated land use transportation planning models: the cases of Sao Paulo, Brazil and Bilbao, Spain. In: Florian M (ed) *The practice of transportation planning*. Elsevier, The Hague, pp 263–286
- Fiorello D, Huismans G, López E, Marques C, Monzon A, Nuijten A, Steenberghen T, Wegener M, Zografos G (2006) Transport strategies under the scarcity of energy supply. STEPs Final report. Buck Consultants International, The Hague
- Hägerstrand T (1970) What about people in regional science? *Pap Reg Sci Assoc* 24(1):7–21
- Hansen WG (1959) How accessibility shapes land use. *J Am Inst Plann* 25(2):73–76

- Harris B, Wilson AG (1978) Equilibrium values and dynamics of attractiveness terms in production-constrained spatial-interaction models. *Environ Plann A* 10(4):371–388
- Hunt JD, Abraham JE (2005) Design and implementation of PECAS: a generalised system for the allocation of economic production, exchange and consumption quantities. In: Lee-Gosselin MEH, Doherty ST (eds) *Integrated land-use and transportation models: behavioural foundations*. Elsevier, St. Louis, pp 253–274
- Hunt JD, Kriger DS, Miller EJ (2005) Current operational urban land-use transport modeling frameworks: a review. *Transport Rev* 25(3):329–376
- Lautso K, Spiekermann K, Wegener M, Sheppard I, Steadman P, Martino A, Domingo R, Gayda S (2004) PROPOLIS: planning and research of policies for land use and transport for increasing urban sustainability. PROPOLIS final report. LT Consultants, Helsinki
- Lee DB (1973) Requiem for large-scale models. *J Am Inst Plann* 39(3):163–178
- Lowry IS (1964) A model of metropolis. RM-4035-RC. Rand Corporation, Santa Monica
- Marshall S, Banister D (eds) (2007) Land use and transport. European research towards integrated policies. Elsevier, London
- Martinez FJ (1996) MUSSA: land use model for Santiago City. *Transportation Res Rec* 1552/1996:126–134
- Miyamoto K, Udomsri R (1996) An analysis system for integrated policy measures regarding land use, transport and the environment in a metropolis. In: Hayashi Y, Roy J (eds) *Transport, land use and the environment*. Kluwer, Dordrecht, pp 259–280
- Nguyen-Luong D (2008) An integrated land-use transport model for the Paris Region (SIMAURIF): ten lessons learned after four years of development. IAURIF, Paris. http://mit.edu/11.521/proj08/readings/D_Mes_documentsDNLpredict3ERSA_2008article_SIMAURIF_10_lessons.pdf. Accessed 24 Mar 2012
- Orcutt G, Greenberger M, Rivlin A, Korbel J (1961) Microanalysis of socioeconomic systems: a simulation study. Harper and Row, New York
- Ravenstein EG (1885) The laws of migration. *J Stat Soc Lond* 48(2):167–235
- Salvini PA, Miller EJ (2005) ILUTE: an operational prototype of a comprehensive microsimulation model of urban systems. *Network Spatial Econ* 5(2):217–234
- Simmonds DC (1999) The design of the DELTA land-use modelling package. *Environ Plann B: Plann Des* 26(5):665–684
- Snickars F, Weibull JW (1977) A minimum information principle. *Reg Sci Urban Econ* 7(1–2):137–168
- von Thünen JH (1826) *Der isolierte Staat in Beziehung auf Landwirtschaft und Nationalökonomie*. Perthes, Hamburg
- Waddell P (2002) UrbanSim: modeling urban development for land use, transportation and environmental planning. *J Am Plann Assoc* 68(3):297–314
- Wagner P, Wegener M (2007) Urban land use, transport and environment models: experiences with an integrated microscopic approach. *disP* 170:3/2007 45–56
- Wegener M (1982) Modeling urban decline: a multilevel economic-demographic model of the Dortmund region. *Int Reg Sci Rev* 7(2):217–241
- Wegener M (2004) Overview of land-use transport models. In: Hensher DA, Button KJ (eds) *Transport geography and spatial systems*. Handbook 5 of handbook in transport. Pergamon/Elsevier Science, Kidlington, pp 127–146
- Wegener M (2011a) The IRPUD model. Arbeitspapier 11/01. Spiekermann & Wegener Stadt- und Regionalforschung, Dortmund
- Wegener M (2011b) From macro to micro – how much micro is too much? *Transport Rev* 31(2):161–177
- Wegener M, Fürst F (1999) Land-use transport interaction: state of the art. Berichte aus dem Institut für Raumplanung 46. Institute of Spatial Planning, University of Dortmund, Dortmund. <http://www.raumplanung.uni-dortmund.de/irpud/fileadmin/irpud/content/documents/publications/ber46.pdf>. Accessed 24 Mar 2012

- Wegener M, Gnad F, Vannahme M (1986) The time scale of urban change. In: Hutchinson B, Batty M (eds) *Advances in urban systems modelling*. North Holland, Amsterdam, pp 145–197
- Wilson AG (1967) A statistical theory of spatial distribution models. *Transportation Res* 1(3):253–269
- Wilson AG (1970) *Entropy in urban and regional modelling*. Pion, London
- Zahavi Y (1974) Traveltime budgets and mobility in urban areas. Report FHW PL-8183. US Department of Transportation, Washington, DC
- Zipf GK (1949) *Human behaviour and the principle of least effort*. Addison Wesley, Cambridge, MA

Network Equilibrium Models for Urban Transport

40

David Boyce

Contents

40.1	Introduction	760
40.2	Historical Overview	761
40.3	Model Formulations	762
40.3.1	Definitions and Assumptions	763
40.3.2	Methodological Approach	764
40.3.3	Deterministic Route Choice over a Road Network	765
40.3.4	Stochastic Route Choice over a Road Network	769
40.3.5	Mode and Route Choice over Road and Fixed Cost Networks	773
40.3.6	O-D, Mode, and Route Choice over Road and Fixed Cost Networks	775
40.4	Model Solution and Implementation	779
40.4.1	Solution Algorithms	779
40.4.2	Unique Route Flows and Multi-class Link Flows	781
40.5	Conclusions	782
	References	785

Abstract

Methods for the analysis and prediction of travel conforming to macroscopic assumptions about choices of the urban population cut a broad swath through the field of regional science: economic behavior, spatial analysis, optimization methods, parameter estimation techniques, computational algorithms, network equilibria, and plan evaluation and analysis. This chapter seeks to expose one approach to the construction of models of urban travel choices and implicitly location choices. Beginning with the simple route choice problem faced by vehicle operators in a congested urban road network, exogenous constants are

D. Boyce

Department of Civil and Environmental Engineering, Northwestern University, Evanston, IL,
USA

e-mail: dboyce@uic.edu

relaxed and replaced with additional assumptions and fewer constants, leading toward a more general forecasting method. The approach, and examples based upon it, reflects the author's research experience of 40 years with the formulation, implementation, and solution of such models.

40.1 Introduction

Journey times and costs are important variables in determining the wide range of choices available to individual travelers. To predict personal travel choices on congested urban road and public transport systems, journey times must be endogenous to the model. This statement is axiomatic. Otherwise, the representation of user congestion, a principal causative agent of urban travel and location choices, is not possible. This axiom provided the foundation for the original formulation of the road traffic network equilibrium model by Martin Beckmann (Beckmann et al. 1956). This seminal contribution, on which the entire field of urban travel choice modeling is implicitly based, was then overlooked for more than a decade. By the time it was rediscovered, a sequential, four-step paradigm had taken hold, consisting of (a) trip generation: the total amount of travel per time period (hour, day) that begins and ends at each location; (b) trip distribution: the amount of travel from every origin to every destination; (c) mode split: the proportion of trips by private cars, trains, buses, cycles, walking, and other modes of travel; and (d) traffic assignment: allocation of modal trip matrices to shortest routes to determine road link and transit line flows. Researchers then began to ask how to combine these steps into a more internally consistent method, only to arrive at Beckmann's original formulation and its extensions. Because of this irony of history, this literature became known as "combined models."

The objective of this chapter is to introduce one type of transportation network user-equilibrium model that originated from Beckmann's formulation: multi-class, multimodal, static models of origin-destination, mode, and route choices. Multi-class refers to models that consider two or more classes of travelers with different behavioral or choice characteristics. Multimodal refers to the consideration of all modes, such as public transport systems, but also including cycling and walking, in addition to motor vehicles on the road network. Static refers to models of constant flows over a congested period, such as the weekday morning or afternoon commuting period, possibly divided into intervals as short as 60 min.

This focus stems from an interest in models that are useful for decision-making about long-range transportation investments as well as short-range demand management. The era of building large-scale urban transportation infrastructure in developed urban economies has largely passed. Now these urban areas are focused on demand management issues, such as road pricing, as well as incremental additions to their road, public transport, and cycle-walkway systems. In contrast, rapidly developing urban economies, especially in Asia, are presently engaged in infrastructure development. Effective and efficient decisions for these systems' investment and management require an advanced evaluation framework to provide

information on the distribution of impacts on residents, employers, and public agencies. Travel forecasts are central to such a framework.

A conviction that travel forecasting models have the potential to be substantially superior to current travel forecasting practice, described by Ortúzar and Willumsen (2011), is one motivation for this chapter. Following a brief historical overview, formulations of several models are offered, beginning with a basic model of route choice on a road network. Assumptions about what is exogenous to that model are then relaxed, enabling a more general model to emerge. Solution algorithms for these models are described, including the issue of the uniqueness of route flows and multi-class link flows. A brief discussion of future research and practice concludes the chapter. References emphasize seminal works in the field and syntheses useful to newcomers.

40.2 Historical Overview

The historical development of this field is complex, in part because separate strands of research and practice provide a variety of approaches. An extensive historical account and mathematically rigorous synthesis of the field with over 1,000 references was prepared by Patriksson (1994). Marcotte and Patriksson (2007) updated and substantially extended that earlier synthesis. Sheffi (1985) synthesized his own contributions on stochastic route choice, as well as integrating some findings of others. Oppenheim (1995) set out to write a textbook on travel demand models, and in addition, offered several theoretical advances to origin-destination-mode-route choice models based on random utility theory. Florian and Hearn (1995) synthesized the network equilibrium literature from the viewpoint of operations research. Bell and Iida (1997) articulated their view of transportation network analysis, including chapters on reliability and design. Nagurney (1999) explored the application of variational inequalities to a variety of network-related problems. A review of implemented combined models was offered by Boyce and Bar-Gera (2004).

This overview is organized by groups of academic researchers working along similar lines. Beckmann did not follow up on his innovation. Instead, research extending Beckmann's model was undertaken by Stella Dafermos and her contemporaries. From her 1968 Ph. D. thesis until her death in 1990, Dafermos established a wide-ranging theory of traffic network equilibrium, including contributions to models with variable and fixed demand, treatment of multiple user classes and asymmetric cost functions, and perhaps most importantly extensions and applications of the theory of variational inequalities to transportation network equilibria. From the late 1970s, Michael Smith independently pursued a similar line of inquiry, focused on traffic equilibrium, traffic signal timing, and road pricing. Patriksson (1994) lists 19 references by Dafermos and 14 references by Smith.

The Centre for Research on Transportation at the University of Montreal, founded in 1972, embarked on theoretical research, model implementation, and testing. Initially led by Michael Florian (2008), successive generations of faculty and students made sustained contributions to network equilibrium modeling. Contributions to solving the transportation network equilibrium problem with variable demand,

including mode choice, were made by Florian and Nguyen during the 1970s. Subsequently, several of these methods were implemented in EMME (www.inro.ca), an interactive-graphic multimodal urban transportation planning system.

In the United Kingdom in the mid-1960s, John Murchland sought to devise an alternative to the sequential paradigm, but it was Suzanne Evans (1976) who devised a way to combine trip distribution and traffic assignment models into a single formulation, an optimization problem consisting of two parts, one related to route choice as in Beckmann's formulation and the other related to trip distribution, as suggested by Wilson (1970). Evans extensively explored the mathematical properties of her formulation and proposed a solution algorithm; see Sect. 40.4.1.

Boyce began to implement the formulation and algorithm of Evans in 1976. Over the next 25 years, he and his students, in separate collaborations with LeBlanc and Lundqvist, implemented a single-class, two-mode combined model on aggregated networks of Chicago and Stockholm. Model parameters were borrowed from other studies at first, but later estimated in a way that is self-consistent with the model solution. Boyce and Bar-Gera (2003) and several collaborators implemented, estimated, and validated a two-class, two-mode combined model at the same level of detail used by transportation planning professionals for the Chicago region.

In 1986, researchers in Chile began to implement multi-class combined models emphasizing route choices in the congested public transport network with several submodes found in Santiago (De Cea et al. 2005). This effort led to the development of ESTRAUS (www.mctsoft.com) which has been applied to Santiago and other Chilean cities. Aashtiani and Magnanti formulated a combined mode choice and traffic assignment model based on nonlinear complementarity theory, and Safwat and Magnanti extended this formulation to include trip generation as well as trip distribution; see Patriksson (1994) for references. Abrahamsson and Lundqvist (1999) extended a model of the Stockholm region to include parameter estimation methods and tested alternative specifications of nested travel choice functions.

The author submits there are different views of how to model urban travel, which are often mutually stimulating to research and practice. For example, another view poses separate travel demand and network cost models, which are solved jointly with an iterative equilibration procedure. From this perspective, there is less emphasis on model integration and more focus on model structure, parameter estimation, and solution procedures for the separate demand and network models. This approach may offer more flexibility to innovative modelers, who indeed often describe themselves as either demand modelers or network modelers, but seldom both. However, it offers fewer opportunities to analyze the properties of the entire model structure and to insure the consistency of the overall approach.

40.3 Model Formulations

Formulations and analyses of combined models of travel choice on congested urban transportation networks based on constrained optimization methods are introduced here, articulating one way to derive models of varying degrees of scope and

complexity. Detailed statements of model properties are omitted, but may be identified using standard techniques for deriving the optimality conditions for a convex function with equality and inequality constraints, as stated in Sect. 40.3.2. The model formulations represent the conventional (traditional) way of describing urban travel, known as trip-based, which originated in the United States in the 1950s. Activity-based or tour-based models, which are more representative of actual travel choices, are the subject of current research and advanced practice, but are not considered here.

40.3.1 Definitions and Assumptions

The following assumptions are briefly stated in agreement with current practice:

1. An urban region is divided into small, relatively homogeneous zones. Zone size varies with the density of development, so that activity levels per zone are relatively similar.
2. Urban activities in zones are described in terms of (a) residential population and households; (b) employment, education (primary, secondary, higher), and day care; (c) shopping, personal and business services, recreation, etc.
3. Facilities for urban activities consist of land and buildings: (a) residences, (b) workplaces, (c) schools, (d) shopping and service centers, and (e) parks and recreational facilities.
4. Travel occurs on two types of transportation systems or modes: (a) private vehicles/ways for driving-cycling-walking/traffic control system and (b) bus or train/roadway or railway/operations plan. Trucks also use the roadway system, depicted in car equivalent units. Transportation systems are represented as networks of nodes, links, link attributes, and for public transport, routes of scheduled services. Service characteristics of links depend on fixed parameters related to physical roadways and vehicle characteristics: (a) length, number, and width of lanes by type, including cycleways and walkways and grade; (b) public transport station spacing and vehicle performance; (c) control and operations plans: speed limits, signal settings, and road tolls; and (d) service frequencies, operating speeds, and public transport fares.
5. Other variables related to travel flows (demand) also determine service characteristics: (a) flows of cars, trucks, cycles, and pedestrians and (b) public transport boardings and alightings at stops per unit time. Taken together, these variables determine the performance characteristics of individual links:

$$\text{link travel time} = f(\text{flows} | \text{fixed vehicle/way characteristics, and operations plans})$$

Such cost performance functions are sometimes confused with supply functions. In a supply function, specific aspects of the vehicle-way-operations plan are not fixed, but are decision variables representing the operator or supplier of services. In contrast, in a travel forecast for a given scenario based on performance functions, optimal values of supply parameters are generally not represented. For example, traffic signal timings and public transport service frequencies are not optimized in response to the travel forecast.

6. Travel between daily activities (residing, working, eating, shopping, schooling, recreating) may be described in terms of pairs of activities linked by trips: (a) homework, (b) work-eat meal, (c) work-shop, (d) shop-home, etc. Over the 24 h weekday, travel related to several activities makes up a sequence of trips connecting various purposes, or tour. The duration of the activities and the times required for travel determine the daily geographic range. In the trip-based approach, individual trips are aggregated by purpose and forecast as separate groups. Whether travel occurs by private car, either alone or with others, by public transport, cycle, or walking, depends on the availability of modes, their relative service times and monetary costs, as well as intangible factors like comfort and convenience. The timing of travel during the day also depends on constraints imposed by activity schedules, and the travel conditions on the private and public networks.
7. Travel occurs during a given period of the 24 h weekday, such as the morning peak commuting period. To represent observed trips, with their specific departure and arrival times, as flows (persons/unit time), a transformation is required, such as (a) all travelers departing from home for work during 6–9 a.m. are counted as flows in persons/hour, and (b) all travelers arriving at work from home during 6–9 a.m. are counted as flows in persons/hour.

For transportation systems planning, facility design, operations planning, or conformance with air quality regulations, forecasts of the following variables are required for each transportation system/activity pattern scenario:

1. Flows of private cars, trucks, cycles, pedestrians, and public transport vehicles on the road network by morning and evening peak commuting periods and for longer periods when travel conditions are relatively stable
2. Flows of persons on the public transport network by submode by time period
3. Flows of persons from origin zone to destination zone by private vehicles and public transport by time period

A capability to examine changes in these flows in response to changes in network layout, capacity and service attributes, monetary costs (e.g., fuel, tolls, fares, parking fees), and changes in zonal activity levels is required.

40.3.2 Methodological Approach

Travel choice models may be formulated and analyzed using several methods for solving optimization and equilibrium problems. These problems include convex optimization, nonlinear complementarity, variational inequality, geometric optimality, and fixed point, roughly in increasing order of generality. Each of these methods has been applied in the formulation of travel choice models. This brief introduction is limited to minimization of a convex function subject to inequality constraints, which is suitable for derivations in this chapter and based on the classic Karush-Kuhn-Tucker theorem (Kuhn and Tucker 1951):

$$\min_{(\mathbf{x})} f(\mathbf{x}) \quad (40.1)$$

$$\text{st: } h_i(\mathbf{x}) \leq 0, \quad i = 1 \dots, m \quad (40.2)$$

where \mathbf{x} is an unknown vector of length n , $f()$ is a strictly convex function, and $h_i(\mathbf{x}) \leq 0$, $i = 1 \dots, m$, is a set of linear constraints. The necessary conditions for $f(\mathbf{x}^*)$ to be a local minimum are

$$\begin{aligned} \frac{\partial f(\mathbf{x}^*)}{\partial x_j} + \sum_{i=1}^m \lambda_i \left(\frac{\partial h_i(\mathbf{x}^*)}{\partial x_j} \right) &\geq 0, \quad j = 1 \dots, n \\ h_i(\mathbf{x}^*) &\leq 0, \quad i = 1 \dots, m \\ \lambda_i h_i(\mathbf{x}^*) &= 0, \quad i = 1 \dots, m \\ \lambda_i &\geq 0, \quad i = 1 \dots, m \end{aligned} \quad (40.3)$$

where λ_i is a dual variable associated with the constraint $h_i(\mathbf{x}^*) \leq 0$. If the inequality constraints include nonnegativity conditions, $\mathbf{x} \geq \mathbf{0}$, then the optimality condition can be written in a more compact and transparent manner, as follows:

$$\frac{\partial f(\mathbf{x}^*)}{\partial x_j} - \sum_{i=1}^m \lambda_i \left(\frac{\partial h_i(\mathbf{x}^*)}{\partial x_j} \right) \geq 0, \quad j = 1 \dots, n \quad (40.4)$$

$$x_j \left(\frac{\partial f(\mathbf{x}^*)}{\partial x_j} - \sum_{i=1}^m \lambda_i \left(\frac{\partial h_i(\mathbf{x}^*)}{\partial x_j} \right) \right) = 0, \quad j = 1 \dots, n \quad (40.5)$$

$$h_i(\mathbf{x}^*) \geq 0, \quad i = 1 \dots, m \quad (40.6)$$

$$\lambda_i h_i(\mathbf{x}^*) = 0, \quad i = 1 \dots, m \quad (40.7)$$

$$x_j \geq 0, \quad j = 1 \dots, n \quad (40.8)$$

$$\lambda_i \geq 0, \quad i = 1 \dots, m \quad (40.9)$$

Equations (40.5) are complementary slackness conditions, which state that either $x_j = 0$, or $\frac{\partial f(\mathbf{x}^*)}{\partial x_j} - \sum_{i=1}^m \lambda_i \left(\frac{\partial h_i(\mathbf{x}^*)}{\partial x_j} \right) = 0$, or both. As shown below, these conditions are needed for deriving the equilibrium conditions on route flows.

40.3.3 Deterministic Route Choice over a Road Network

In 1952, John Wardrop, a British traffic scientist, proposed the following criterion to describe traffic flows on a route:

The journey times on all routes actually used are equal, and less than those which would be experienced by a single vehicle on any unused route. . . . (this) criterion is quite a likely one

in practice, since it might be assumed that traffic will tend to settle down into an equilibrium situation in which no driver can reduce his journey time by choosing a new route. (Patriksson 1994, p. 31)

The first sentence is now known as Wardrop's first principle of network user equilibrium (UE).

Beckmann et al. (1956, p. 59), working at the Cowles Commission for Research in Economics at the University of Chicago during 1951–1954, described the concept of equilibrium more generally:

Demand refers to trips and capacity refers to flows on roads. The connecting link is found in the distribution of trips over the network according to the principle that traffic follows shortest routes in terms of average cost. The idea of equilibrium in a network can then be described as follows. The prevailing demand for transportation, that is, the existing pattern of originations and terminations, gives rise to traffic conditions that will maintain that same demand. Or, starting at the other end, the existing traffic conditions are such as to call forth the demand that will sustain the flows that create these conditions.

Then, they described their concept of route equilibrium as follows (p. 60):

... the principle of traffic distribution among alternative routes in equilibrium. (1) If between a given origin and a given destination more than one route is actually traveled, the cost of transportation to the average road user, as indicated by the average-cost capacity curves, must be equal on all these routes. (2) Since the routes used are the "shortest" ones under prevailing traffic conditions, average cost on all other possible routes cannot be less than that on the route or routes traveled. (3) The amount of traffic originated per unit of time must equal the demand for transportation at the trip cost which prevails.

Note that these statements reflect Beckmann's view that origin–destination demand is variable, whereas Wardrop considered fixed flows (p. 344). McGuire stated they were not aware of Wardrop's paper at the time, although he became aware of it later (personal interview, 1999).

By "routes actually used," Wardrop meant the routes used from a given origin to a given destination, which may be defined as zones. If routes consist of sequences of links, then route costs may be defined as the sum of the link costs along the route. Since link costs depend on link flows, through the cost performance function, and each link serves (possibly) many routes, then identifying the route costs which satisfy Wardrop's principle involves solving the route choice problem simultaneously for a system of zones and a network.

Link flows have units of vehicles/hour (vph), so route flows and origin–destination (O-D) flows also have units of vehicles/hour or persons/hour. The resulting route choice model is a steady-state flow model in which no individual travels from an origin to a destination. Rather, O-D-route flows occur with corresponding flows on the links of each used route. This formulation leads in a relatively simple concept of congestion, with no bottlenecks or traffic jams, but only steadily flowing vehicles traveling at speeds determined by cost performance functions.

The user-equilibrium link flows and costs, and a set of route flows, corresponding to fixed O-D flows may be determined by solving the following constrained optimization problem:

$$\begin{aligned}
\min_{(\mathbf{h})} z(\mathbf{h}) &= \sum_{a \in A} \int_0^{f_a} c_a(x) dx \\
\text{st: } &\sum_{r \in R_{pq}} h_r = \bar{d}_{pq}, \quad p \in P; q \in Q \\
&h_r \geq 0, \quad r \in R_{pq}, p \in P; q \in Q
\end{aligned} \tag{40.10}$$

$$\text{where } f_a \equiv \sum_{pq} \sum_{r \in R_{pq}} h_r \delta_{ar}, a \in A$$

f_a = flow of all vehicles on link a (vph)

$c_a(f_a)$ = generalized travel cost function for link a , a nondecreasing function of link flow f_a

h_r = flow of vehicles on route r of the set of routes R_{pq} connecting zone p to zone q (vph)

\bar{d}_{pq} = exogenous flow of vehicles from zone p to zone q (vph)

$\delta_{ar} = 1$, if link a belongs to route r from zone p to zone q and 0 otherwise

P, Q = sets of origin and destination zones, respectively

A = set of links in the network

The unknown variables are vehicle route flows $\mathbf{h} = (h_r)$; vehicle link flows (f_a) are defined in terms of the route flows. The link-route correspondence matrix (δ_{ar}) is exogenous. To simplify the derivation, truck flows are included in the single O-D matrix (\bar{d}_{pq}).

The optimality conditions for the above problem may be stated as follows:

$$\begin{aligned}
\sum_{a \in A} c_a(f_a) \delta_{ar} - u_{pq} &\geq 0, \quad r \in R_{pq}, \quad p \in P, q \in Q \\
h_r \left(\sum_{a \in A} c_a(f_a) \delta_{ar} - u_{pq} \right) &= 0, \quad r \in R_{pq}, \quad p \in P, q \in Q \\
\left(\sum_{r \in R_{pq}} h_r - \bar{d}_{pq} \right) &\geq 0, \quad p \in P, q \in Q \\
u_{pq} \left(\sum_{r \in R_{pq}} h_r - \bar{d}_{pq} \right) &= 0, \quad p \in P, q \in Q \\
h_r \geq 0, \quad r \in R_{pq}, \quad u_{pq} \geq 0, \quad p \in P, q \in Q
\end{aligned} \tag{40.11}$$

where u_{pq} is a dual variable associated with the conservation of flow constraint defined on the exogenous O-D flow \bar{d}_{pq} . Conditions (40.11) may be interpreted as follows for O-D pair pq :

1. Assume $h_r > 0$; then, $\left(\sum_{a \in A} c_a(f_a) \delta_{ar} - u_{pq} \right) = 0$, or $C_r \equiv \sum_{a \in A} c_a(f_a) \delta_{ar} = u_{pq}$

Table 40.1 Deterministic models

Choice	Equilibrium conditions
Route	$h_r > 0 \Rightarrow C_r = u_{pq}; h_s = 0 \Rightarrow C_s \geq u_{pq}; C_t > u_{pq} \Rightarrow h_t = 0;$ $C_r \equiv \sum_{a \in A} c_a(f_a) \delta_{ar}, r \in R_{pq}, \sum_{r \in R_{pq}} h_r = \bar{d}_{pq}, p \in P, q \in Q$
Mode and route	$h_r^c > 0 \Rightarrow C_r^c = u_{pq}^c; h_r^c = 0 \Rightarrow C_r^c \geq u_{pq}^c; \sum_{r \in R_{pq}} h_r^c = d_{pq}^c$ $C_r^c > u_{pq}^c \Rightarrow h_r^c = 0; C_r^c \equiv \sum_{a \in A} c_a(f_a) \delta_{ar}, r \in R_{pq}^c$ $d_{pq}^c > 0 \Rightarrow u_{pq}^c = \kappa_{pq}; d_{pq}^c = 0 \Rightarrow u_{pq}^c \geq \kappa_{pq}$ $u_{pq}^c > \kappa_{pq} \Rightarrow d_{pq}^c = 0$ $d_{pq}^n > 0 \Rightarrow C_{pq}^n = \kappa_{pq}; d_{pq}^n = 0 \Rightarrow C_{pq}^n \geq \kappa_{pq}$ $C_{pq}^n > \kappa_{pq} \Rightarrow d_{pq}^n = 0, n \in N$
O-D, mode, and route	Deterministic equilibrium conditions for O-D flows correspond to the solution of a cost-minimizing allocation of origins to destinations, known as the classical transportation problem of linear programming (Evans 1973). The solution corresponds to a deterministic model for $\eta \rightarrow \infty$ in the case of the O-D-mode model and the mode-O-D model. Based on empirical studies of origin-destination flows, such solutions are considered to be unrealistic for urban travel choices
Mode, O-D, and route	

2. Assume $h_s = 0$; then, $\left(\sum_{a \in A} c_a(f_a) \delta_{as} - u_{pq} \right) \geq 0$, or $C_s \geq u_{pq}$.

3. Assume $C_t > u_{pq}$; then $h_t = 0$.

where C_r is the travel cost on route r , the sum of the costs of the links comprising route r . Hence, every used route connecting zone p to zone q has a generalized travel cost equal to u_{pq} , and no unused route has a lower travel cost. Thus, this formulation corresponds to Wardrop's first principle. These conditions are shown in Table 40.1 in the first row.

In the above formulation, identical and rational travelers are assumed to know accurately their travel times over alternative routes from their origins to destinations. The source of this information can only be described as being from their experience. Modern in-vehicle navigation systems may offer travel time information over one route at some point in time, but generally not over several alternative routes. Moreover, the model formulation applies to a relatively long time period, such as the morning peak period, during which actual route travel times may vary widely. This assumption implies that the model solution corresponds to a deterministic user equilibrium (DUE) and is relaxed somewhat in the next subsection.

If the generalized cost functions are strictly increasing with link flows, then the objective function is strictly convex guaranteeing that the solution is unique in the link flows. The solution is not unique in the route flows, or class link flows in the case that two or more demand classes are specified, however, since the total link flows are linear functions of the route flows. Therefore, the objective function is not strictly convex in the route flows, as required for uniqueness. The above

formulation applies to the case in which each link performance function depends only on the link's own flow, which is called separable. In a more general case, called symmetric, each link performance function depends on a specified vector of link flows such that the effect of a change in link a 's flow on link b equals the effect of a change in link b 's flow on link a for all links specified in the cost function. An example of such a vector of link flows is the links entering an intersection. Generally, intersection delays are not symmetric, so this requirement is not met. Models not exhibiting such symmetries are called asymmetric and may be formulated variational inequality problems (Patriksson 1994, pp. 74–77).

In contrast to the above model, the conventional approach to modeling route choice on a public transport network is relatively simple:

1. Represent all submodes (bus, rapid transit, commuter rail, express bus) in one network.
2. Find a minimal generalized travel cost route from origin zone p to destination zone q considering access, waiting, boarding, in-vehicle, and transfer times as well as fares; congestion at boarding and alighting are not considered.
3. Assign all public transport trips from zone p to zone q to a single minimal cost route, which is called all-or-nothing assignment.

Even if all-or-nothing assignment to minimal cost routes is considered adequate, representing public transport networks is more complex than road networks. The use of only one route between each O-D pair is simplistic if several public transport options are offered. Methods for modeling public transport route choice are found in Ortúzar and Willumsen (2011, pp. 373–80).

40.3.4 Stochastic Route Choice over a Road Network

A way to relax the deterministic route choice model, and possibly make it more realistic, is to introduce an additional constraint. The role of this constraint is to soften or blur the deterministic character of the above model by allowing some portion of each O-D flow to choose routes with higher costs. Before exploring this idea, it is appropriate to ask how many routes are used by each O-D pair in the deterministic solution. To answer this question, a moderately congested car O-D matrix was computed for the 1790 zone system of the Chicago region, and added to a truck O-D matrix obtained from the region's planning organization; see Bar-Gera and Boyce (2007). The total vehicle flow of 1,349,000 vph between 3,174,000 zone pairs was assigned to the Chicago regional network. The total number of UE routes in a very precise solution was 8,573,000, or 2.70 routes per O-D pair.

Of these O-D pairs, about 55 % have only one route, which seems surprising since many of these routes are very long. About 90 % of O-D pairs have five or fewer routes, and 99 % have 20 or fewer routes. However, one O-D pair has 1920 routes, the maximum in this solution. Figure 40.1 shows the number of O-D pairs on the y-axis versus the number of routes per O-D pair on the x-axis. The cumulative number of O-D pairs is shown starting at 1.0 at the upper left, decreasing to 1E-7 (0.0000001) at the lower right. Note where the line crosses the second horizontal

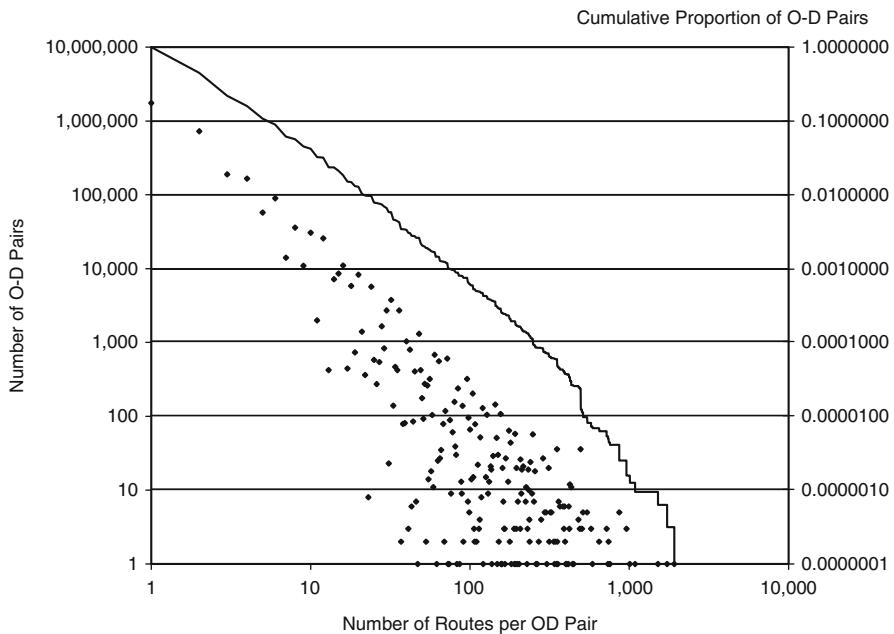


Fig. 40.1 Number of O-D pairs versus number of routes per O-D pair

line labeled 0.10 on the right y-axis; the five dots to the left of this intersection account for 90 % of all O-D pairs. In a more congested solution, the number of routes is much larger.

If one wished to distribute some O-D flow to higher cost routes, how might this be accomplished? Such a redistribution may be considered to be a “dispersion” of choices from the cost-minimizing UE routes to higher cost routes. A function depicting such a dispersion of proportions of route flow is available for this purpose. Known as the entropy function (Erlander and Stewart 1990, pp. 21–25), it has a one-to-one correspondence with the well-known logit function.

A constraint can be formed to represent such a dispersion. Travelers strictly take the least cost route in the deterministic solution, so it is the least dispersed feasible solution to problem (40.10). By constraining the route choices to be greater than this minimum level, some choices are shifted to higher cost routes. The form of the constraint is

$$-\sum_{p \in P} \sum_{q \in Q} \sum_{r \in R_{pq}} h_r \ln(h_r) > S_{UE} \quad (40.12)$$

where S_{UE} represents the dispersion of the choices in the DUE solution. Since there are unlikely to be any data at present on the dispersion of routes in a large network, a route dispersion constraint is simply a conceptual device. Modifying conditions (40.11) to include the effect of the dispersion constraint, one may obtain

$$\begin{aligned}
& \sum_{a \in A} c_a(f_a) \delta_{ar} + \frac{1}{\theta} (\ln h_r + 1) - u_{pq} \geq 0, \quad r \in R_{pq}, \quad p \in P, q \in Q \\
& h_r \left(\sum_{a \in A} c_a(f_a) \delta_{ar} + \frac{1}{\theta} (\ln h_r + 1) - u_{pq} \right) = 0, \quad r \in R_{pq}, \quad p \in P, q \in Q \\
& \left(\sum_{r \in R_{pq}} h_r - \bar{d}_{pq} \right) \geq 0, \quad p \in P, q \in Q \\
& u_{pq} \left(\sum_{r \in R_{pq}} h_r - \bar{d}_{pq} \right) = 0, \quad p \in P, q \in Q \\
& S + \sum_{p \in P} \sum_{q \in Q} \sum_{r \in R_{pq}} h_r \ln(h_r) \geq 0 \\
& \frac{1}{\theta} \left(S + \sum_{p \in P} \sum_{q \in Q} \sum_{r \in R_{pq}} h_r \ln(h_r) \right) = 0 \\
& h_r \geq 0, \quad r \in R_{pq}, \quad u_{pq} \geq 0, \quad p \in P, q \in Q, \quad \frac{1}{\theta} > 0
\end{aligned} \tag{40.13}$$

where $S > S_{UE}$ is the dispersion of an entropy-constrained solution, and $1/\theta$ is the dual variable corresponding to the entropy constraint. The reason that it is defined as a reciprocal will become clear shortly. Because the route flow h_r appears as the argument of the natural logarithm, it cannot take on a value of zero. Hence, all route flows are positive. Solving the complementarity slackness condition for $h_r > 0$, the following optimality conditions may be obtained:

$$\begin{aligned}
& \ln h_r = \theta (u_{pq} - C_r) - 1, \quad \text{so} \\
& h_r = \exp(\theta u_{pq} - 1 - \theta C_r), \quad \text{where } C_r \equiv \sum_{a \in A} c_a(f_a) \delta_{ar}
\end{aligned} \tag{40.14}$$

Apply the conservation of route flow constraint to this expression for h_r to obtain

$$\sum_{r \in R_{pq}} h_r = \bar{d}_{pq} = \exp(\theta u_{pq} - 1) \sum_{r \in R_{pq}} \exp(-\theta C_r) \tag{40.15}$$

$$\exp(\theta u_{pq} - 1) = \frac{\bar{d}_{pq}}{\sum_{r \in R_{pq}} \exp(-\theta C_r)} \tag{40.16}$$

Substituting this expression into the equation for h_r yields the logit route choice function:

$$h_r = \bar{d}_{pq} \frac{\exp(-\theta C_r)}{\sum_{r \in R_{pq}} \exp(-\theta C_r)} \tag{40.17}$$

These conditions are shown in the first row of [Table 40.2](#). Examination of these conditions reveals the structure of the stochastic user-equilibrium (SUE) model as well as raising several issues. Corresponding to the dispersion constraint, a logit

Table 40.2 Stochastic and mixed stochastic-deterministic models

Choices	Functions and equilibrium conditions
Route	<p>Stochastic route choice:</p> $C_r \leq C_r^{\max} \Rightarrow h_r > 0; C_r > C_r^{\max} \Rightarrow h_r = 0$ $h_r = \bar{d}_{pq} \frac{\exp(-\theta C_r)}{\sum_{r \in R_{pq}} \exp(-\theta C_r)}, \quad \theta < \infty$
Mode and route	<p>Stochastic mode choice with deterministic route choice:</p> $\theta \rightarrow \infty \Rightarrow C_{pq}^c \equiv u_{pq}^c$ $d_{pq}^m = \bar{d}_{pq} \frac{\exp(-\mu C_{pq}^m)}{\sum_{m \in M} \exp(-\mu C_{pq}^m)}, \quad m \in M$ <p>Stochastic mode choice with stochastic route choice:</p> $\theta < \infty \Rightarrow \tilde{C}_{pq}^c \equiv -\frac{1}{\theta} \ln \sum_{r \in R_{pq}} \exp(-\theta C_r)$ $h_r^c = \bar{d}_{pq} \frac{\exp(-\mu \tilde{C}_{pq}^c)}{\sum_{m \in M} \exp(-\mu C_{pq}^m)} \frac{\exp(-\theta C_r)}{\sum_{r \in R_{pq}} \exp(-\theta C_r)}, \quad \mu < \theta < \infty$
O-D, mode, and route	$\mu < \infty \Rightarrow \tilde{C}_{pq} = -\frac{1}{\mu} \sum_{m \in M} \exp(-\mu C_{pq}^m)$ $d_{pq} = A_p \bar{O}_p B_q \bar{D}_q \exp(-\eta \tilde{C}_{pq}), \text{ where}$ $A_p \Rightarrow \sum_{q \in Q} \sum_{m \in M} d_{pq}^m = \bar{O}_p; B_q \Rightarrow \sum_{p \in P} \sum_{m \in M} d_{pq}^m = \bar{D}_q$ $d_{pq}^m = A_p \bar{O}_p B_q \bar{D}_q \exp(-\eta \tilde{C}_{pq}) \frac{\exp(-\mu C_{pq}^m)}{\sum_{m \in M} \exp(-\mu C_{pq}^m)}, \quad \eta < \mu < \infty$
Mode, O-D, and route	$\eta < \infty \Rightarrow \hat{C}_p^m = -\frac{1}{\eta} \ln \sum_{q \in Q} D_q B_q \exp(-\eta C_{pq}^m)$ $d_{pq}^m = d_p^m \frac{B_q \bar{D}_q \exp(-\eta C_{pq}^m)}{\sum_q B_q \bar{D}_q \exp(-\eta C_{pq}^m)}, \text{ where } B_q \Rightarrow \sum_{p \in P} \sum_{m \in M} d_{pq}^m = D_q$ $d_{pq}^m = \bar{O}_p \frac{\exp(-\mu \hat{C}_p^m)}{\sum_{m \in M} \exp(-\mu C_p^m)} \frac{B_q \bar{D}_q \exp(-\eta C_{pq}^m)}{\sum_q B_q \bar{D}_q \exp(-\eta C_{pq}^m)}, \quad \mu < \eta < \infty$

route choice function is obtained. The logit function allocates some portion of the O-D flow \bar{d}_{pq} to every route connecting p to q . Of course, as $C_r \Rightarrow +\infty$, $\exp(-\theta C_r) \Rightarrow 0$. For small values of θ , including 0, this limit could be ineffective. Therefore, a limit may need to be placed on the definition of routes and the maximum value of C_r . One possibility is that in traversing each link of a route, a driver must travel farther in cost from the origin and closer to the destination. However, the equilibrium travel costs are not available for the purpose of defining these route costs, so a fixed link cost, such as the free-flow time, might be used. Another option is to place an upper limit on the travel cost for each O-D pair, which could be related to the deterministic travel cost: $C_r^{SUE} \leq C_r^{\max} \equiv k C_r^{DUE}, r \in R_{pq}, k > 1$.

Another problem with the logit route choice function is that the SUE routes are often not distinct alternatives. Rather, routes may have substantial overlapping segments in an urban application. As Patriksson (1994, p. 65) noted, O-D flows are over allocated to overlapping routes. Consequently, the logit model may yield overly large flows. This anomaly results from the inability of the logit model to account for the correlation between the costs of alternative routes, which stems from its independence from irrelevant alternatives property. Another problem is that the route choice probabilities are based solely on route cost differences, and do not take their magnitudes into account. For additional discussion of SUE models, see Patriksson (1994, pp. 60–65) and Sheffi (1985).

40.3.5 Mode and Route Choice over Road and Fixed Cost Networks

If travel choices are viewed as a hierarchy, then route choice would seem to fall naturally at the lowest level. Route choices are specific to each transportation mode and may vary with every trip in response to current travel conditions. The choice of mode may be hypothesized to be the next higher choice in this hierarchy. Mode choices may be made on a daily basis, assuming a car is available. Choice of cycle and walking may also be considered, if available as options. Note the representation of choices as a hierarchy does not imply that choices are made sequentially.

To begin, the deterministic car route choice model is extended to include choice of mode to serve as a bridge to the consideration of stochastic models of mode as well as route choice. Denote mode by the superscript $m = \{c, n\}$ where c represents the car mode and $n \in N$ represents the subset of other modes (public transport, cycle, and walk), assumed to have fixed travel costs and not to affect the congestion of the car mode. The addition of superscript c to the demand for O-D pair pq denotes person flows per hour from zone p to zone q by car; likewise, n denotes person flows per hour by other modes. Route flows by car h_r^c are now redefined as persons per hour, whereas link flows f_a are defined as vehicles per hour, the conversion from persons to vehicles occurring in the definitional equation for link flows.

Consider the following generalization of the deterministic car route choice problem:

$$\begin{aligned}
 \min_{(\mathbf{h}^c, \mathbf{d})} z(\mathbf{h}^c, \mathbf{d}) &= v \sum_{a \in A} \int_0^{f_a} c_a(x) dx + \sum_{pq} \sum_{n \in N} C_{pq}^n d_{pq}^n \\
 \text{st: } \sum_{r \in R_{pq}^c} h_r^c &= d_{pq}^c, \quad p \in P; q \in Q \\
 \sum_{m \in M} d_{pq}^m &= \bar{d}_{pq}, \quad p \in P; q \in Q \\
 h_r^c &\geq 0, \quad r \in R_{pq}^c \quad p \in P; q \in Q \\
 d_{pq}^m &\geq 0, \quad m \in M, \quad p \in P; q \in Q \\
 \text{where } f_a &\equiv \sum_{pq} \sum_{r \in R_{pq}^c} h_r^c \delta_{ar} / v, \quad a \in A
 \end{aligned} \tag{40.18}$$

C_{pq}^n = fixed generalized cost of travel per person by fixed cost mode n from zone p to zone q

d_{pq}^c = flow of persons from zone p to zone q by the car mode c

d_{pq}^n = flow of persons from zone p to zone q by a fixed cost mode n , and

$$\mathbf{d} = \begin{pmatrix} d_{pq}^m \\ d_{pq}^n \end{pmatrix}$$

v = mean car occupancy, the ratio of all car occupants to all cars (persons per vehicle)

Note the following additions and changes in model formulation (40.18):

1. A parameter v is inserted in the first term of the objective function, and the definition of link flows, enabling the units of persons and vehicles to be represented consistently.
2. A second term is added to the objective function defining the total costs of fixed cost modes.
3. A new constraint requires the O-D-mode flows to sum to the exogenous O-D flow \bar{d}_{pq} .
4. The set of routes by car is redefined as R_{pq}^c .

This formulation seeks to find the allocation of exogenous O-D flows to the several modes, and within the car mode to routes, so as to minimize a function of total generalized cost. Because car route flows should be UE, the total cost of car travel is not minimized; rather, the sum of the integrals of the links cost functions is retained. The derivation of the optimality conditions with respect to h_r^c , d_{pq}^c , and d_{pq}^n is left as an exercise for the reader; their interpretation is presented in the second row of Table 40.1, where u_{pq}^c is a dual variable associated with the car route flow conservation constraints, and κ_{pq} is a new dual variable associated with the O-D flow conservation constraints.

To interpret these conditions, consider the case of $h_r^c > 0$, for an O-D pair pq . The UE route travel cost is $C_r^c \equiv \sum_{a \in A} c_a(f_a) \delta_{ar} = u_{pq}^c$, as is true for all used car routes from zone p to zone q . If $h_r^c > 0$, then $d_{pq}^c > 0$, and $u_{pq}^c = \kappa_{pq}$, the equilibrium modal O-D cost from zone p to zone q . If $\kappa_{pq} = C_{pq}^n$, for one or more of the fixed cost modes $n \in N$, then $d_{pq}^n \geq 0$; otherwise, $C_{pq}^n > \kappa_{pq}$, and $d_{pq}^n = 0$. The following conclusions may be drawn for this deterministic formulation:

1. If O-D flows occur by car, then all used routes have equal cost, and no unused route has a lower cost.
2. The UE costs of the used car routes not only determine the O-D cost but also determine whether any of the fixed cost modes (public transport, cycle, and walk) have sufficiently low costs to be used: if $u_{pq}^c < C_{pq}^n$, then $d_{pq}^n = 0$ (no one uses mode n). If $u_{pq}^c = C_{pq}^n$, then the O-D cost of fixed cost mode n and car are equal, and use of mode n may occur. If $u_{pq}^c > C_{pq}^n = \kappa_{pq}$, then all O-D flow occurs by one or more fixed cost modes, such as public transport from an outer suburb to the CBD, and no one uses car. That is, either there is no fixed cost mode flow or the fixed mode cost sets a maximum level for the car costs for each O-D pair. Hence, the solution is “all-or-nothing” with respect to mode.

3. If the car occupancy v were not added to the first term of the objective function, and to the definition of link flow, then the car O-D cost would be different from the O-D equilibrium cost by a factor equal to v . For consistency of the formulation, then, the parameter v is needed in the objective function.

One often observes travel by two or more modes (car, public transport, cycle, or walk) between many O-D pairs in survey data. Therefore, the formulation of the mode and car route choice model as a deterministic cost minimization problem, while instructive, is unrealistic. The relaxation of this deterministic formulation is proposed through the addition of a modal dispersion constraint, as in the stochastic route choice model. A function representing modal dispersion may be imposed to make mode choices more dispersed than the DUE minimum level; that is, some choices are allocated to higher cost modes. The form of the constraint is

$$-\sum_{pq} \sum_{m \in M} d_{pq}^m \ln(d_{pq}^m) \geq S \quad (40.19)$$

where S represents the level of dispersion of the choices to higher cost modes. Note: S cannot be observed except in very simple cases in which all of the observed choices are enumerated. S cannot be determined from sample data because a sample by its nature is less dispersed (more clustered) than the population. Let this constraint be added to the mode and route problem (40.18) above. The analysis of the optimality conditions is shown in the Mode and Route row of [Table 40.2](#).

Let $1/\mu$ be the dual variable associated with the dispersion constraint. Then, in the same way as the logit route choice function was derived in [Sect. 40.3.4](#), a logit mode choice function may be derived, as shown in the upper panel of the Mode and Route row of [Table 40.2](#). This choice function includes the fixed cost modes (public transport, cycle, and walk), as well as car with its endogenous deterministic route costs. Together with the same UE conditions for car route costs, the function depicts the equilibrium conditions for stochastic mode and deterministic route choice. By replacing the car deterministic route conditions with the stochastic route choice function, a combined stochastic mode and route choice function may be obtained, as shown in the lower panel of the Mode and Route row of [Table 40.2](#). Here the O-D cost of the car mode is the “composite cost” derived from the denominator of the logit route choice function. This composite cost replaces the equal route costs of each O-D pair governed by the deterministic conditions. The derivation of such a composite cost is given in the next section.

40.3.6 O-D, Mode, and Route Choice over Road and Fixed Cost Networks

The combined mode and route choice formulation can be further extended to include an origin–destination dispersion function in the same manner as described above for mode choice. In the version presented here, constraints are added to the mode and route choice formulation to derive a model corresponding to the classical

trip distribution function (Wilson 1970, pp. 15–17). These constraints consist of origin and destination constraints and another dispersion function representing dispersion of trips to higher cost destinations, separately from the modal dispersion constraint. In the following development, the relationship of these two constraints is explored. This formulation may be stated as follows, by further augmenting problem (40.18):

$$\begin{aligned}
 \min_{(\mathbf{h}^c, \mathbf{d})} z(\mathbf{h}^c, \mathbf{d}) &= v \sum_{a \in A} \int_0^{f_a} c_a(x) dx + \sum_{pq} \sum_{n \in N} C_{pq}^n d_{pq}^n \\
 \text{st: } &\sum_{r \in R_{pq}^c} h_r^c = d_{pq}^c, \quad p \in P, q \in Q \\
 &\sum_m d_{pq}^m = d_{pq}, \quad p \in P, q \in Q \\
 &-\sum_{pqm} d_{pq}^m \ln\left(\frac{d_{pq}^m}{d_{pq}}\right) \geq S^M \\
 &\sum_q d_{pq} = \bar{O}_p \quad p \in P \\
 &\sum_p d_{pq} = \bar{D}_q \quad q \in Q \\
 &-\sum_{pq} d_{pq} \ln(d_{pq}) \geq S^{PQ} \\
 &h_r^c \geq 0, \quad r \in R_{pq}^c, p \in P, q \in Q \\
 &d_{pq}^m > 0, \quad m \in M, p \in P, q \in Q \\
 &d_{pq} > 0, \quad p \in P, q \in Q, \\
 \text{where } f_a &\equiv \sum_{pq} \sum_{r \in R_{pq}^c} h_r^c \delta_{ar}/v, \quad a \in A
 \end{aligned} \tag{40.20}$$

The O-D-mode flow d_{pq}^m is assumed to be conditional on the O-D flow d_{pq} through its insertion into the denominator of the mode dispersion constraint as an a priori flow. The modal flows d_{pq}^m are constrained to sum to the O-D flow by the mode conservation of flow constraint. The origin–destination dispersion constraint is defined on S^{PQ} , and the O-D flows are constrained by the exogenous origin and destination totals, \bar{O}_p and \bar{D}_q .

Analysis of the UE conditions for car proceeds in the same way as in the mode and route choice models with regard to UE car cost C_{pq}^c . Consider the optimality conditions for d_{pq}^m and d_{pq} :

$$\begin{aligned}
 \ln\left(\frac{d_{pq}^m}{d_{pq}}\right) &= \mu\left(\kappa_{pq} - C_{pq}^m\right) - 1 \\
 \ln(d_{pq}) &= \eta\left(\alpha_p + \beta_q - \kappa_{pq} - \frac{1}{\mu}\right) - 1
 \end{aligned} \tag{40.21}$$

where $1/\eta$ is the dual variable for the O-D dispersion constraint and α_p and β_q are respectively the dual variables for the origin and destination constraints. Solving the first condition for d_{pq}^m , and applying the mode conservation of flow constraint, yields

$$d_{pq}^m = d_{pq} \exp(\mu \kappa_{pq} - 1) \exp(-\mu C_{pq}^m) \quad (40.22)$$

$$\sum_{m \in M} d_{pq}^m = d_{pq} = d_{pq} \exp(\mu \kappa_{pq} - 1) \sum_{m \in M} \exp(-\mu C_{pq}^m) \quad (40.23)$$

Solving for the exponential function containing κ_{pq} ,

$$\exp(\mu \kappa_{pq} - 1) = 1 / \sum_{m \in M} \exp(-\mu C_{pq}^m) \quad (40.24)$$

which can then be substituted into Eq. (40.22) to yield for the case of the car mode:

$$d_{pq}^c = d_{pq} \frac{\exp(-\mu C_{pq}^c)}{\sum_{m \in M} \exp(-\mu C_{pq}^m)} \quad (40.25)$$

This result expresses the O-D-car flow as the O-D flow times a logit function based on the UE cost for car C_{pq}^c and the costs of the fixed costs modes C_{pq}^m . A similar expression may be derived for the fixed cost modes. Now define $\exp(-\mu \tilde{C}_{pq}) \equiv \sum_{m \in M} \exp(-\mu C_{pq}^m)$; taking logs and solving for \tilde{C}_{pq} gives the modal composite cost from zone p to zone q ,

$$\tilde{C}_{pq} = -\frac{1}{\mu} \ln \sum_{n \in M} \exp(-\mu c_{pqn}) \quad (40.26)$$

Note that Eq. (40.24) can be rearranged as $\exp(-\mu(\kappa_{pq} + \frac{1}{\mu})) = \sum_{m \in M} \exp(-\mu C_{pq}^m)$. Therefore,

$$\tilde{C}_{pq} = \left(\kappa_{pq} + \frac{1}{\mu} \right) \quad (40.27)$$

An expression for the O-D flow d_{pq} can then be derived from optimality condition (40.21):

$$d_{pq} = \exp(\alpha_p + \beta_q - 1) \exp\left(-\eta\left(\kappa_{pq} + \frac{1}{\mu}\right)\right) = \exp(\alpha_p + \beta_q - 1) \exp(-\eta \tilde{C}_{pq}) \quad (40.28)$$

By applying the origin and destination constraints, a more compact expression may be obtained:

$$d_{pq} = A_p \bar{O}_p B_q \bar{D}_q \exp(-\eta \tilde{C}_{pq}) = \frac{\bar{O}_p \bar{D}_q \exp(-\eta C_{pq})}{\sum_q B_q \bar{D}_q \exp(-\eta C_{pq}) \sum_p A_p \bar{O}_p \exp(-\eta C_{pq})} \quad (40.29)$$

where A_p and B_q are balancing factors defined by Eq. (40.29) that insure that \bar{O}_p , the exogenous originating flow from zone p , and \bar{D}_q , the exogenous terminating flow at zone q , are satisfied (Wilson 1970, pp. 22–25). By substituting the O-D function for d_{pq} into the O-D-mode function for d_{pq}^m , the combined O-D-mode function may be stated as

$$d_{pq}^m = A_p \bar{O}_p B_q \bar{D}_q \exp(-\eta \tilde{C}_{pq}) \frac{\exp(-\mu C_{pq}^m)}{\sum_{m \in M} \exp(-\mu C_{pq}^m)} \quad (40.30)$$

This model may be extended further to include stochastic route choice, as described in Sect. 40.3.4. According to one hierarchy hypothesized to motivate the dispersion constraints, route choices are deterministic cost-minimizing functions of car travel costs, mode choices tend to be cost minimizing with some dispersion to the higher cost mode, and O-D choices are even less cost minimizing. By this rationale, the cost sensitivity parameters estimated from survey data for the logit functions should have numerical values such that $\eta \leq \mu$; a larger value of the parameter means that travelers are more sensitive to the fixed transport costs and UE car costs than for a smaller parameter value. For an interpretation of these coefficients based on the utilities of the choices, see Williams (1977, pp. 330–336) and Oppenheim (1995, pp. 198–205).

These values have additional implications in the logit function context. The cross elasticities of flow (demand) with respect to mode choice may be negative if $\eta > \mu$, meaning that an increase in the cost c' of a mode m' would lead to a decrease in the demand for traffic on the competing mode m'' , contradicting what intuitively would be expected from the transportation system (Abrahamsson and Lundqvist 1999, p. 93). An implication of estimated parameter values that violate this condition is that the hypothesis of the model is incorrect and that mode choice is less cost minimizing than O-D choice or, equivalently, that O-D choice should be conditional on mode choice. This situation led Abrahamsson and Lundqvist (1999, pp. 86–87) to hypothesize the “reverse nested combined model,” shown in the fourth row of Table 40.2. For this hypothesis, mode choice is less cost minimizing than O-D choice. If the cost of the modes are very different (low for car and high for public transport), a very small value of μ could be required for the estimated model to predict the sample choices correctly.

Boyce and Bar-Gera (2003) found that the parameter size condition was violated for other travel (not home-to-work travel) during the morning peak period for the Chicago area in 1990. Formulation and solution of such models for forecasting and scenario analysis is only meaningful if travel is segmented into several homogeneous classes, and implemented for time periods during the day with relatively stable levels of congestion.

40.4 Model Solution and Implementation

The solution and implementation of combined models of urban travel choice proceeded slowly in comparison with the sequential procedure in travel forecasting practice. Even so, combined models have provided a framework and basis for evaluating solution methods used in practice. This section briefly traces the evolution of solution methods for combined models, and describes a few notable efforts regarding their implementation and validation, concluding with a discussion of a new traffic assignment method for finding unique route flows and multi-class link flows.

40.4.1 Solution Algorithms

When the formulation of a model of variable demand and route choice on a network was first proposed by Beckmann et al. (1956), no solution algorithm was offered. Despite the needs of transportation planning studies in the United States, 1955–1975, and the United Kingdom, 1960–1975, to forecast urban travel for congested conditions, the potential contribution of Beckmann's formulation was not recognized. By the late 1960s, several Ph.D. students had rediscovered Beckmann's formulation and began to propose solution algorithms. Among these, the algorithm of Suzanne Evans was the most detailed and promising (Evans 1976). Evans proposed an iterative, convergent algorithm for trip distribution and traffic assignment (O-D and road route choice) that linearized the objective function only as necessary, and otherwise used the O-D choice functions directly. Her algorithm may be summarized as follows:

Step 1 Find an initial solution: for free-flow link travel costs by road, find the least cost car routes between all zone pairs, compute an initial car O-D matrix with the travel choice function, and assign it to the least cost routes, resulting in an initial link flow vector; these arrays define a current feasible solution.

Step 2 For the travel costs corresponding to the current road link flow vector, find the least cost car routes, compute a new O-D matrix with the travel choice function, and assign the car O-D matrix to the least cost routes, resulting in a new link flow vector.

Step 3 Find weights $(1 - \lambda)$ and λ , $0 \leq \lambda \leq 1$, when used to compute a weighted average of the current and new O-D matrices and link flow vectors, minimizes

the objective function of the formulation augmented by the nonlinear dispersion function times its dual variable.

Step 4 Compute a convergence measure for the updated O-D matrix and link flow vector; if the solution has not converged to the target level, update the current solution and return to step 2; otherwise, stop.

The above algorithm is a partial linearization method for solving a convex optimization problem (Patriksson 1994, pp. 104–111). Although convergent, and useful for solutions on mainframe computers of the 1970–1980s era, the method converges slowly after the first several iterations. At that time, computer resources were generally insufficient to permit more than a few iterations for implementations of several hundred zones and a few thousand links typical of that period. A related algorithm, now known as the Frank-Wolfe method, began to be used from the mid-1970s to solve the road traffic assignment problem with fixed O-D flows.

Combined models of travel choices were implemented and estimated since the early 1980s. To be realistic for practice, such models should represent two or more classes of travelers plus trucks. An early multi-class model was implemented and estimated by Lam and Huang (1992). A model implementation similar in scale to those used in practice was undertaken by Boyce and Bar-Gera (2003) for the morning peak period of the Chicago region. Extensive estimation studies were undertaken; the model was validated with travel-to-work data from the 1990 US Census, contributing new methods for model validation as well as model estimation. That model was solved with the Evans algorithm, the state of the art at that time.

As computers expanded in size and speed during the 1980s with the introduction of supercomputers, engineering workstations and personal computers of similar memory and speed, algorithms for solving the traffic assignment problem with more precision and speed were proposed (Bar-Gera 2002; Dial 2006). These algorithms were origin-based, in contrast to the link-based assignment algorithm based on the full linearization of the objective function. Bar-Gera and Boyce (2003) applied Bar-Gera's origin-based traffic assignment algorithm to devise a solution method for the origin-destination, mode, and car route choice problem that achieved more precise convergence than is possible with the Evans algorithm for large-scale problems. This algorithm replaced the link-based assignment in step 2 with an origin-based procedure to update the solution of the assignment problem. Then the O-D matrices are updated followed by another assignment update, continuing until the convergence criterion is met. Unlike the Evans algorithm, a line search and averaging of solutions are not required.

De Cea et al. (2005) implemented and estimated a combined model for Santiago, Chile. A software system for solving the model, ESTRAUS, was created and applied in the redesign of the public transport system of Santiago (<http://en.wikipedia.org/wiki/Transantiago>). STGO, a software application in EMME, was created to implement a closely related model (Florian et al. 2002). These two systems represent two further implementations of combined models that are used by practitioners. CUBE (www.citilabs.com), TransCAD (www.caliper.com), and VISUM (www.ptv.de) also have the possibility to serve as platforms for

implementing combined models. However, such implementations require substantial knowledge and programming skills.

Solution of the stochastic route choice problem for large networks at a level of precision similar to DUE remains a work in progress. Lee et al. (2010) provided a detailed literature review, proposed two new algorithms, and reported computational results for a problem of moderate size. The use of a method proposed by Bell to find routes avoids the use of a maximum route length. Other problems remain, however, including overlapping routes, as discussed in Sect. 40.3.4.

40.4.2 Unique Route Flows and Multi-class Link Flows

For project evaluation and scenario analyses, total link flows and class O-D flows may suffice. More detailed analyses, however, require O-D-route flows or class link flows. Neither is uniquely determined by the solution of the standard traffic assignment formulation. Computed route flows and class link flows may be quite arbitrary. A simple example in Bar-Gera et al. (2012) illustrates this dilemma, which is well known to researchers and advanced practitioners.

To choose among the infinite possibilities of route flow solutions for a UE model, an additional behavioral assumption is required. One plausible assumption is proportionality, namely, that the proportion of O-D flows assigned to each of two alternative route segments with precisely equal costs should be the same regardless of their origin or their destination. Proportionality also determines class link flows uniquely in multi-class assignments.

The fixed demand traffic assignment with proportionality can be solved in two ways. First, the standard assignment problem can be solved with an origin-based algorithm to a precise level of convergence, such as a relative gap equal to 1E-7. Then the route flows can be adjusted with a post-processing procedure to achieve the same proportions for each O-D flow over each pair of alternative segments, leaving the link flows unchanged. This procedure is now available in the TransCAD and VISUM software systems. Second, the proportionality condition can be used to design a new algorithm to solve the assignment problem. This approach was the basis for TAPAS (Bar-Gera 2010). Comparisons of solutions with TAPAS versus link-based and route-based tools for the Chicago regional road network were presented in Bar-Gera et al. (2012).

An example of route flows over a pair of alternative segments in the road network of the Chicago region is considered next. Two O-D matrices representing cars and trucks were assigned with TAPAS to the Chicago network by imposing the user-equilibrium principle with proportionality. The total flow of vehicles per hour in the matrices is 984,717 cars and 445,185 trucks in car equivalent units. The matrices were assigned to two networks: (a) an unrestricted network in which trucks can use any link and (b) a restricted network in which trucks are prohibited from using 563 car-only links (car-only lanes of two freeways, the Lake Shore Drive, and boulevards and other roads with truck prohibitions). According to the proportionality condition, class O-D flows using a pair of alternative segments should have the same proportion

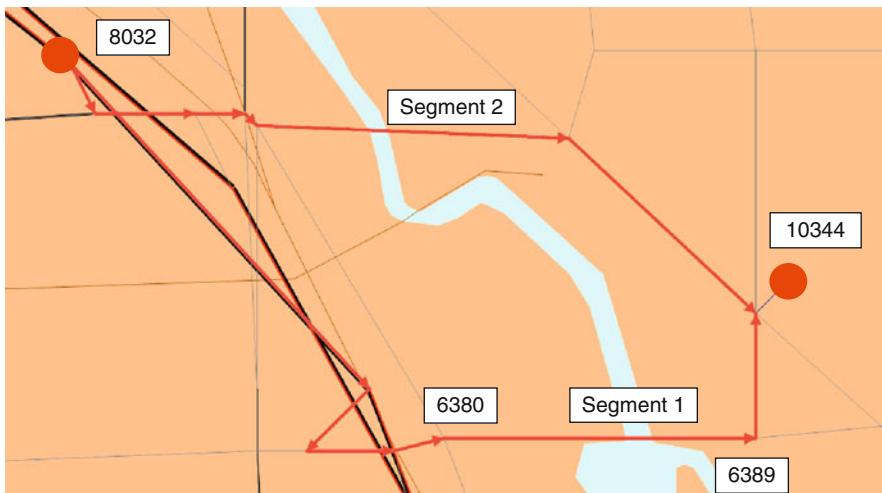


Fig. 40.2 Pair of alternative segments in the Chicago road network

on each segment for each assignment. Since the generalized cost variable is defined to be travel time, the same proportions should be observed for cars and trucks over a pair of segments with no truck restrictions. A pair of segments connecting nodes 8032 and 10344, shown in Fig. 40.2, was selected for this example.

Figure 40.3 compares the total O-D-segment flows on Segment 2 (y-axis) with the total O-D-segment flows on Segment 1 (x-axis). These flows lie on a straight line, showing that the condition of proportionality is imposed. The slopes of the lines are slightly different, indicating only a small change in the proportions between the solutions for the restricted and unrestricted networks. Although the alternative segments have precisely equal travel times in each of the two solutions, the total flows are somewhat different, as shown at the top of the figure for the two solutions. Figure 4 shows the car flows for segments 1 and 2. Note the slopes in Fig. 40.4 are the same as in Fig. 40.3. The truck flows, the differences between the flows in Figs. 40.3 and 40.4, are not shown, but have the same slopes.

The application of the condition of proportionality provides a meaningful and practical solution to the problem of nonuniqueness of route flows and multi-class link flows. TAPAS offers a method for rapidly and precisely solving the traffic assignment problem with proportionality.

40.5 Conclusions

Despite 60 years of research on urban transportation network equilibrium, many problems remain unsolved, and practice lags increasingly behind research knowledge. Research problems may be broadly classified according to travel choices, network representation, network design, and solution procedures, among others.

Fig. 40.3 Total route flows over two networks: (a) 683 segment pairs on unrestricted network (*squares*); (b) 678 segment pairs on restricted network (*triangles*)

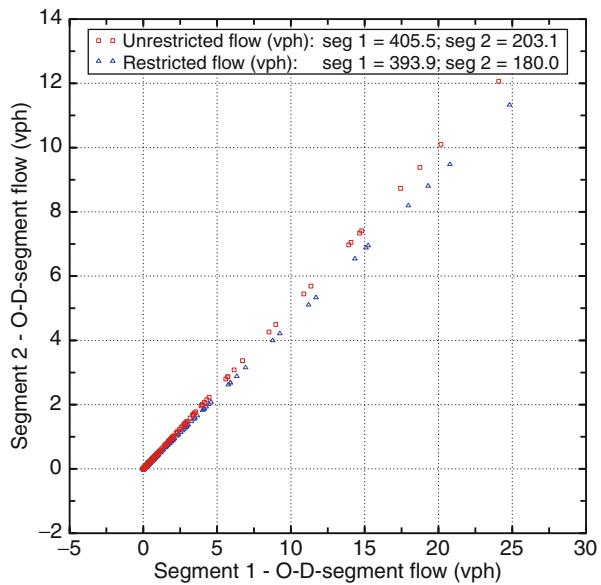
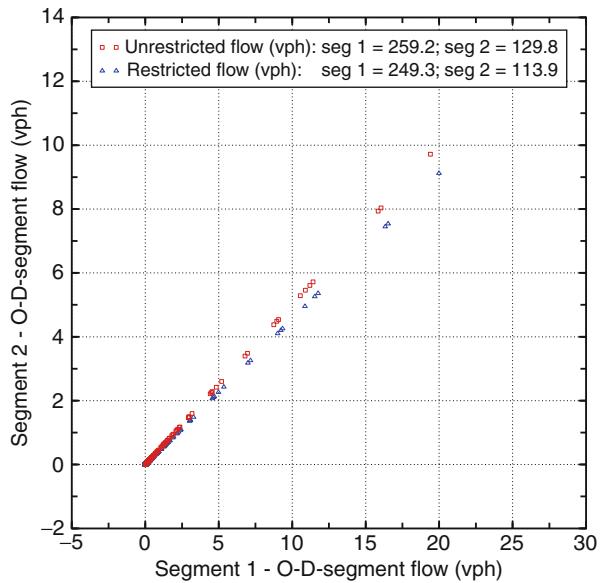


Fig. 40.4 Car route flows over two networks: (a) 680 segment pairs on unrestricted network (*squares*); (b) 670 segment pairs on restricted network (*triangles*)



- Until now, the modeling of travel choices has mainly followed the trip-based paradigm. Increasingly, travel demand modelers view travel in terms of daily tours or daily activities. Generally, prediction of tours or activities has been approached from a micro-simulation point of view. To the author's knowledge,

- the stability of such simulations has generally not been examined, such as by generating a sufficiently large sample of simulations and analyzing their variation. Model formulations of aggregate tour-based travel using concepts similar to those described in this chapter have been studied, but more investigation is needed (Bernardin et al. 2009).
- b. The representation of travel cost functions in network equilibrium models has advanced little beyond the original separable formulation of Beckmann. Although asymmetric models can be formulated as variational inequalities, the apparent lack of uniqueness of solutions has discouraged serious efforts to investigate this approach further. For example, the side-constrained method of Patriksson (1994, pp. 66–70) has not been investigated with large networks. Solution methods have generally not advanced beyond the so-called diagonalization (relaxation) method (Marcotte and Patriksson 2007, pp. 671–673).
 - c. Another problem that has received very little attention is transportation network design. In a combinatorial sense, the network design problem is intractable because of its large size. Other approaches are possible, however, such as the spacing of freeways in a grid, as was considered early in the history of this field. A new approach to this problem might be to develop methods to generate and evaluate scenarios in a systematic and semiautomated manner. Such a method would require an ability to distinguish among the merits of closely related scenarios. Given the precision of solutions to network models now possible, this capability may now be achievable.
 - d. Although practitioners are required by government regulations in the USA to solve their sequential travel forecasting procedures in a way that achieves an internal consistency of travel costs, there is no general agreement on how this should be done. No practitioner or software developer has described, tested, and demonstrated that one procedure is best among alternatives for complex models applied in practice. Moreover, the errors introduced in forecasts that do not achieve consistency remain unknown. This relatively straightforward research problem should be tackled.
 - e. Academic interest in the solution of combined model formulations of travel choice has influenced travel forecasting practice, but only to a limited extent. Except for Santiago, Chile, combined models have rarely been applied in practice. The formulation of a combined model clearly enhances the understanding of the challenges facing the practitioner in solving the model sequence. Few practitioners, however, seem equipped by their training or mathematical ability to gain from insights from these formulation. Moreover, software developers have not incorporated tools in their software systems to facilitate the application of this approach. Based upon past experience, they will not do so until interest among practitioners strongly induces them to proceed.
- Pursuit of this research agenda requires a knowledge of optimization methods, computer skills, data, and perseverance. Many similar problems could be identified, especially from the viewpoint of other perspectives. For those so inclined, the journey will be challenging, but always interesting, and hopefully rewarding.

Acknowledgments Professor Huw Williams, Cardiff University, offered many useful comments on earlier drafts of this chapter. Dr. Hillel Bar-Gera, Ben-Gurion University of the Negev, has offered many stimulating insights and contributions to my thinking on combined network equilibrium models during the past 15 years. Dr. Yu (Marco) Nie, Northwestern University, has been a stimulating colleague during my renewed association with my undergraduate alma mater.

Their contributions are greatly appreciated. Remaining errors are my responsibility.

References

- Abrahamsson T, Lundqvist L (1999) Formulation and estimation of combined network equilibrium models with applications to Stockholm. *Transport Sci* 33(1):80–100
- Bar-Gera H (2002) Origin-based algorithm for the traffic assignment problem. *Transport Sci* 36(4):398–417
- Bar-Gera H (2010) Traffic assignment by paired alternative segments. *Transport Res B* 44(8–9):1022–1046
- Bar-Gera H, Boyce D (2003) Origin-based algorithms for combined travel forecasting models. *Transport Res B* 37(5):405–422
- Bar-Gera H, Boyce D (2007) Some amazing properties of road traffic network equilibria. In: Friesz TL (ed) *Network science, nonlinear science and infrastructure systems*. Springer, Berlin, pp 305–335
- Bar-Gera H, Boyce D, Nie Y (2012) User-equilibrium route flows and the condition of proportionality. *Transport Res B* 46(3):440–462
- Beckmann M, McGuire CB, Winsten CB (1956) *Studies in the economics of transportation*. Yale University Press, New Haven
- Bell MGH, Iida Y (1997) *Transportation network analysis*. Wiley, Chichester
- Bernardin VL Jr, Koppelman F, Boyce D (2009) Enhanced destination choice models incorporating agglomeration related to trip chaining while controlling for spatial competition. *Transport Res Rec* 2131:143–151
- Boyce D, Bar-Gera H (2003) Validation of urban travel forecasting models combining origin–destination, mode and route choices. *J Regional Science* 43(3):517–540
- Boyce D, Bar-Gera H (2004) Multiclass combined models for urban travel forecasting. *Netw Spat Econ* 4(1):115–124
- De Cea J, Fernandez JE, Soto A, Dekock V (2005) Solving network equilibrium on multimodal urban transportation networks with multiple user classes. *Transport Rev* 25(3):293–317
- Dial RB (2006) A path-based user-equilibrium traffic assignment algorithm that obviates path storage and enumeration. *Transport Res B* 40(10):917–936
- Erlander S, Stewart NF (1990) The gravity model in transportation analysis. VSP, Utrecht
- Evans SP (1973) A relationship between the gravity model for trip distribution and the transportation problem in linear programming. *Transport Res* 7(1):39–61
- Evans SP (1976) Derivation and analysis of some models for combining trip distribution and assignment. *Transport Res* 10(1):37–57
- Florian M (2008) Models and software for urban and regional transportation planning: contributions of the center for research on transportation. *INFOR* 46(1):29–49
- Florian M, Hearn D (1995) Network equilibrium models and algorithms. In: Ball MO, Magnanti TL, Monma CL, Nemhauser GL (eds) *Network routing, handbooks in operations research and management science* 8. Elsevier Science, Amsterdam, pp 485–550
- Florian M, Wu JH, He S (2002) A multi-class multi-mode variable demand network equilibrium model with hierarchical logit structures. In: Gendreau M, Marcotte P (eds) *Transportation and network analysis: current trends*. Kluwer, Dordrecht, pp 119–133
- Kuhn HW, Tucker AW (1951) Nonlinear programming. In: Neyman J (ed) *Proceedings of the second Berkeley symposium on mathematical statistics and probability*. University of California Press, Berkeley, pp 481–492

- Lam WHK, Huang H-J (1992) A combined trip distribution and assignment model for multiple user classes. *Transport Res B* 26(4):275–287
- Lee D-H, Meng Q, Deng W (2010) Origin-based partial linearization method of the stochastic user equilibrium traffic assignment problem. *J Transp Eng-ASCE* 136:52–60
- Marcotte P, Patriksson M (2007) Traffic equilibrium. In: Barnhart C, Laporte G (eds) *Transportation, handbooks in operations research and management science* 14. Elsevier Science, Amsterdam, pp 623–713
- Nagurney A (1999) *Network economics*, 2nd edn. Kluwer, Boston
- Oppenheim N (1995) *Urban travel demand modeling*. Wiley, New York
- Ortúzar JD, Willumsen LG (2011) *Modelling transport*, 4th edn. Wiley, New York
- Patriksson M (1994) The traffic assignment problem: models and methods. VSP, Utrecht
- Sheffi Y (1985) *Urban transportation networks*. Prentice-Hall, Englewood Cliffs
- Williams HCWL (1977) On the formation of travel demand models and economic evaluation measures of user benefit. *Environ Plann* 9(3):285–344
- Wilson AG (1970) *Entropy in urban and regional modeling*. Pion, London

Anna Nagurney

Contents

41.1	Introduction	788
41.2	Fundamental Decision-Making Concepts and Models	792
41.2.1	The User-Optimized Problem	793
41.2.2	The System-Optimized Problem	797
41.2.3	The Braess Paradox	799
41.3	Models with Asymmetric Link Costs	801
41.3.1	Variational Inequality Formulations of Fixed Demand Problems	802
41.3.2	Variational Inequality Formulations of Elastic Demand Problems	805
41.4	Conclusions	808
	References	809

Abstract

We overview some of the major advances in supply chains and transportation networks, with a focus on their common theoretical frameworks and underlying behavioral principles. We emphasize that the foundations of supply chains as network systems can be found in the regional science and spatial economics literature. In addition, transportation network concepts, models, and accompanying methodologies have enabled the advancement of supply chain network models from a system-wide and holistic perspective.

We discuss how the concepts of system optimization and user optimization have underpinned transportation network models and how they have evolved to enable the formulation of supply chain network problems operating (and

A. Nagurney

Department of Finance and Operations Management, Isenberg School of Management, University of Massachusetts, Amherst, MA, USA
e-mail: nagurney@isenberg.umass.edu

managed) under centralized or decentralized, that is, competitive, decision-making behavior.

We highlighted some of the principal methodologies, including variational inequality theory, that have enabled the development of advanced transportation network equilibrium models as well as supply chain network equilibrium models.

41.1 Introduction

Supply chains are networks of suppliers, manufacturers, transportation service providers, storage facility managers, retailers, and consumers at the demand markets. Supply chains are the backbones of our globalized network economy and provide the infrastructure for the production, storage, and distribution of goods and associated services as varied as food products, pharmaceuticals, vehicles, computers, and other high-tech equipment, building materials, furniture, clothing, toys, and even electricity.

Supply chains may operate (and be managed) in a centralized or decentralized manner and be underpinned not only by multimodal transportation and logistical networks but also by telecommunication as well as financial networks. In a centralized supply chain, there is a central entity or decision-maker, such as a firm, that controls the various supply chain network activities, whereas in a decentralized supply chain, there are multiple economic decision-makers, and the governing paradigm is that of competitive behavior among the relevant stakeholders, with different degrees of cooperation. For example, in a vertically integrated supply chain, the same firm may be responsible for production, storage, and distribution of its products. On the other hand, certain industry supply chain network structures may consist of competitive manufacturers, competitive distributors, as well as competing retailers. Nevertheless, the stakeholders involved in supply chains must cooperate to the extent that the products be received and processed as they move downstream in the supply chain (Nagurney 2006).

The complexity and interconnectivity of some of today's product supply chains have been vividly illustrated through the effects of recent natural disasters, including earthquakes, tsunamis, and even hurricanes, which have severed critical nodes and/or links and have disrupted the production and transportation of products, with major economic implications. Indeed, when supply chain disruptions occur, whether due to natural disasters, human error, attacks, or even market failure, the ramifications can propagate and impact the health and well-being of the citizenry thousands of miles away from the initially affected location (cf. Nagurney and Qiang 2009).

Since supply chains are network systems, any formalism that seeks to model supply chains and to provide quantifiable insights and measures must be a system-wide one and network based. Such crucial issues as the stability and resiliency of supply chains, as well as their adaptability and responsiveness to events in a global

environment of increasing risk and uncertainty, can only be rigorously examined from the view of supply chains as network systems (Nagurney 2006).

Supply chains share many of the same characteristics as other network systems, including a large-scale nature and complexity of network topology; congestion; which leads to nonlinearities; alternative behavior of users of the networks, which may lead to paradoxical phenomena (recall the well-known Braess paradox in which the addition of a new road may increase the travel time for all); possibly conflicting criteria associated with optimization (the minimization of time for delivery, e.g., may result in higher emissions); interactions among the underlying networks themselves, such as the Internet with electric power networks, financial networks, and transportation and logistical networks; and the growing recognition of their fragility and vulnerability. Moreover, policies surrounding supply chain networks today may have major implications not only economically but also socially, politically, and security-wise.

Although, historically, supply chain activities of manufacturing, transportation/distribution, as well as inventorying/storage have each, independently, received a lot of attention from both researchers and practitioners, the framework of supply chains views the various activities of production, transportation, and consumption in an integrated, holistic manner. Indeed, without the critical transportation links, what is manufactured cannot be delivered to points of demand. Moreover, needed inputs into the production processes/manufacturing links cannot be secured.

While, beginning in the 1980s (cf. Handfield and Nichols 1999), supply chains have captured wide interest among practitioners as well as researchers, it may be argued that the foundations of supply chain networks can be found in regional science and spatial economics, dating to the classical spatial price equilibrium models of Samuelson (1952) and Takayama and Judge (1971) with additional insights as to production processes, transportation, and distribution provided by Beckmann et al. (1956). For example, in spatial price equilibrium models, not only is production of the commodity in question considered at multiple locations or supply markets, with appropriate underlying functions, but also the consumption of the commodity at the demand markets, subject to appropriate functions (either demand or demand price) as well as the cost associated with transporting the commodity between pairs of the spatially separated supply and demand markets. Spatial price equilibrium models have evolved to include multiple commodities and multiple modes of transportation and may even include general underlying transportation networks. Moreover, with advances in theoretical frameworks, including, for example, the theory of variational inequalities (Nagurney 1999), one can now formulate and solve complex spatial price equilibrium problems with asymmetric supply price, demand price, and unit transportation/transaction cost functions (for which an optimization reformulation of the governing spatial price equilibrium conditions does not hold).

In addition, versions of spatial equilibrium models that capture oligopolistic behavior under imperfect, as opposed to perfect, competition serve as some of the basic supply chain network models in which competition is included, but, at the

same time, the important demand/consumption side is also captured (see Nagurney (1999) and the references therein).

Interestingly, spatial price equilibrium problems can be reformulated and solved as transportation network equilibrium problems with elastic demands over appropriately constructed abstract networks or supernetworks (see Nagurney and Dong 2002). Hence, the plethora of algorithms that have been developed for transportation networks (cf. Sheffi 1985; Patriksson 1994; Nagurney 1999; Ran and Boyce 1996) can also be applied to compute solutions to spatial price equilibrium problems. It is worth noting that Beckmann, McGuire, and Winsten in their classical 1956 book, *Studies in the Economics of Transportation*, formulated transportation network equilibrium problems with elastic demands. They proved that under the assumed user link cost functional forms and the travel disutility functional forms associated with the origin/destination pairs of nodes that the governing equilibrium conditions (now known as user-optimized conditions) in which no traveler has any incentive to alter his route of travel, given that the behavior of others is fixed, could be reformulated and solved as an associated optimization problem. In their book, they also hypothesized that electric power generation and distribution networks, or in today's terminology, electric power supply chains, could be transformed into transportation network equilibrium problems. This has now been established (cf. Nagurney (2006) and the references therein).

Today, the behavior of travelers on transportation networks is assumed to follow one of Wardrop's (1952) two principles of travel behavior, now renamed, according to Dafermos and Sparrow (1969), as user-optimized (selfish or decentralized) or system-optimized (unselfish or centralized). The former concept captures individuals' route-taking decision-making behavior, whereas the latter assumes a central controller that routes the flow on the network so as to minimize the total cost.

Moreover, a plethora of supply chain network equilibrium models, originated by Nagurney et al. (2002), have been developed in order to address competition among decision-makers in a tier of a supply chain whether among the manufacturers, the distributors, the retailers, and/or even the consumers at the demand markets. Such models capture the behavior of the individual economic decision-makers, as in the case, for example, of profit maximization, and acknowledge that consumers also take transaction/transportation costs into consideration in making their purchasing decisions. Prices for the product associated with each decision-maker at each tier are obtained once the entire supply chain network equilibrium problem is solved, yielding also the equilibrium flows of the product on the links of the supply chain network. Such supply chain network equilibrium models also possess (as spatial price equilibrium problems highlighted above) a transportation network equilibrium reformulation.

Supply chain network models have been generalized to include electronic commerce options, multiple products, as well as risk and uncertainty on the demand-side as well as on the supply side (cf. Nagurney (2006) and the referenced therein). In addition, and, this is product-specific, supply chain network models have also been constructed to handle time-sensitive products (fast fashion, holiday based, and even critical needs as in disasters) as well as perishable products (such as

food, cut flowers, certain vaccines and medicines, etc.) using multicriteria decision-making formalisms for the former and generalized networks for the latter (see Masoumi et al. 2012). Both static as well as dynamic supply chain network models, including multiperiod ones with inventorying, have been formulated, solved, and applied.

It is important to note that not all supply chains are commercial, and, in fact, given that the number of disasters is growing, as is the number of people affected by them, humanitarian supply chains have emerged as essential elements in disaster recovery. Unlike commercial or corporate supply chains, humanitarian supply chains are not managed using profit maximization as a decision-making criterion (since donors, e.g., would not approve), but rather cost minimization subject to demand satisfaction under uncertainty is relevant (see Nagurney and Qiang 2009). In addition, such supply chains may need to be constructed quickly and with the cognizant decision-makers working under conditions of damaged, if not destroyed, infrastructure and limited information.

Supply chain decision-making occurs at different levels – at the strategic, tactical, and operational levels. Strategic decisions may involve where to locate manufacturing facilities and distribution centers, whereas tactical decisions may include with which suppliers to partner and which transportation service providers (carriers) to use. Decisions associated with operational supply chain decision-making would involve how much of the product to produce at which manufacturing plants, which storage facilities to use and how much to store where, as well as how much of the product should be supplied to the different retailers or points of demand. In addition, because of globalization, supply chain decision-making may now involve outsourcing decisions as well as the accompanying risk management.

Today, it has been argued that, increasingly, in the network economy it is not only competition within a product supply chain that is taking place but, rather, supply chain versus supply chain competition. Zhang et al. (2003) generalized Wardrop's first principle of travel behavior to formulate competition among supply chains.

Location-based decisions are fundamental to supply chain decision-making, design, and management. Furthermore, such decisions affect spatial competition as well as trade, with Ohlin (1933) and Isard (1954) noting the need to integrate industrial location and international trade in a common framework.

Nagurney (2010) constructed a system-optimization model that can be applied to the design or redesign of a supply chain network and has as endogenous variables both the capacities associated with the links (corresponding to manufacturing, transportation, and storage) as well as the operational flows of the product in order to meet the demands. The model has been extended in various directions to handle oligopolistic competition as well as product perishability in specific applications (cf. Masoumi et al. (2012) and the references therein).

At the same time that supply chains have become increasingly globalized, environmental concerns due to global warming and associated risks have drawn the attention of numerous constituencies. Firms are increasingly being held accountable not only for their own performance in terms of their environmental performance but also for that of their suppliers, subcontractors, joint venture

partners, distribution outlets, and, ultimately, even for the disposal of their products. Consequently, poor environmental performance at any stage of the supply chain may damage the most important asset that a company has, which is its reputation. Hence, the topic of sustainable supply chain network modeling and analysis has emerged as an essential area for research, practice, as well as for policy analysis (see Boone et al. 2012).

41.2 Fundamental Decision-Making Concepts and Models

In this section of this chapter, we interweave fundamental concepts in transportation that have been used successfully and with wide application in supply chain network modeling, analysis, operations management, and design. Our goal is to provide the necessary background from which additional explorations and advances can be made using a readable and accessible format.

As noted in the introduction, over half a century ago, Wardrop (1952) considered alternative possible behaviors of users of transportation networks, notably, urban transportation networks, and stated two principles, which are named after him:

First principle: The journey times of all routes actually used are equal and less than those which would be experienced by a single vehicle on any unused route.

Second principle: The average journey time is minimal.

The first principle corresponds to the behavioral principle in which travelers seek to (unilaterally) determine their minimal costs of travel; the second principle corresponds to the behavioral principle in which the total cost in the network is minimal.

Beckmann et al. (1956) were the first to rigorously formulate these conditions mathematically and proved the equivalence between the *transportation network equilibrium* conditions, which state that all used paths connecting an origin/destination (O/D) pair will have equal and minimal travel times (or costs) (corresponding to Wardrop's first principle), and the Kuhn-Tucker conditions of an appropriately constructed optimization problem, under a symmetry assumption on the underlying functions. Hence, in this case, the equilibrium link and path flows could be obtained as the solution of a mathematical programming problem. Their fundamental result made the formulation, analysis, and subsequent computation of solutions to transportation network problems based on actual transportation networks realizable.

Dafermos and Sparrow (1969) coined the terms *user-optimized* (U-O) and *system-optimized* (S-O) transportation networks to distinguish between two distinct situations in which, respectively, travelers act unilaterally, in their own self-interest, in selecting their routes and in which travelers choose routes/paths according to what is optimal from a societal point of view, in that the total cost in the network system is minimized. In the latter problem, marginal total costs rather than average costs are equilibrated. As noted in the introduction, the former problem coincides with Wardrop's first principle and the latter with Wardrop's second principle. Table 41.1 highlights the two distinct behavioral principles underlying transportation networks.

Table 41.1 Distinct behavior on transportation networks

<i>User optimization</i>	<i>System optimization</i>
↓	↓
<i>User equilibrium principle</i>	<i>System-optimality principle</i>
User travel costs on used paths for each O/D pair are equalized and minimal	Marginals of the total travel cost on used paths for each O/D pair are equalized and minimal

The concept of “system optimization” is also relevant to other types of “routing models” in transportation, including those concerned with the routing of freight. Dafermos and Sparrow (1969) also provided explicit computational procedures, that is, *algorithms*, to compute the solutions to such network problems in the case where the user travel cost on a link was an increasing (in order to handle congestion) function of the flow on the particular link and linear. Today, the concepts of user optimization versus system optimization also capture, respectively, decentralized versus centralized decision-making on supply chain networks after the proper identifications are made (Boyce et al. 2005; Nagurney 2006).

In this section, the basic transportation network models are first recalled, under distinct assumptions as to their operation and the underlying behavior of the users of the network. The models are classical and are due to Beckmann et al. (1956) and Dafermos and Sparrow (1969). In subsequent sections, we present more general models in which the user link cost functions are no longer separable but, rather, are asymmetric. For such models, we also provide the variational inequality formulations of the governing equilibrium conditions, since, in such cases, the governing equilibrium conditions can no longer be reformulated as the Kuhn-Tucker conditions of a convex optimization problem. The presentation follows that in Nagurney (2007) with addition of material on supply chains with synthesis.

For easy accessibility, we recall the classical user-optimized network model in Sect. 41.2.1 and then the classical system-optimized network model in Sect. 41.2.2. The Braess (1968) paradox is, subsequently, highlighted in Sect. 41.2.3.

41.2.1 The User-Optimized Problem

The user-optimized network problem is also commonly referred to in the transportation literature as the *traffic assignment* problem or the *traffic network equilibrium* problem.

Consider a general network $\mathcal{G} = [\mathcal{N}, \mathcal{L}]$, where \mathcal{N} denotes the set of nodes and \mathcal{L} the set of directed links. Links connect pairs of nodes in the network and are denoted by a, b , etc. Let p denote a path consisting of a sequence of links connecting an origin/destination (O/D) pair of nodes. Paths are assumed to be acyclic and are denoted by p, q , etc. In transportation networks, nodes correspond to origins and destinations, as well as to intersections. Links, on the other hand, correspond to

roads/streets in the case of urban transportation networks and to railroad segments in the case of train networks. A path in its most basic setting, thus, is a sequence of “roads” which comprise a route from an origin to a destination. In the supply chain network context, links correspond to supply chain activities (with appropriate associated cost functions) and represent manufacturing, transportation/shipment, storage, etc. In addition, links can correspond to outsourcing links (see Nagurney 2006).

Here we consider *paths*, rather than *routes*, since the former subsumes the latter. The network concepts presented here are sufficiently general to abstract not only transportation decision-making but also combined/integrated location-transportation decision-making as well as a spectrum of supply chain decisions. In addition, in the setting of *supernetworks*, that is, abstract networks, in which nodes need to correspond to locations in space (see Nagurney and Dong 2002), a path is viewed more broadly and need not be limited to a route-type decision but may, in fact, correspond to not only transportation but also to manufacturing and inventorying/storage decision-making.

Let P_ω denote the set of paths connecting the origin/destination (O/D) pair of nodes ω . Let P denote the set of all paths in the network and assume that there are J origin/destination pairs of nodes in the set Ω . Let x_p represent the nonnegative flow on path p and let f_a denote the flow on link a . All vectors here are assumed to be column vectors. The path flows on the network are grouped into the vector $x \in R_+^{n_P}$, where n_P denotes the number of paths in the network. The link flows, in turn, are grouped into the vector $f \in R_+^{n_L}$, where n_L denotes the number of links in the network.

Assume, as given, the demand associated with each O/D pair ω , which is denoted by d_ω , for $\omega \in \Omega$. In the network, the following conservation of flow equations must hold:

$$d_\omega = \sum_{p \in P_\omega} x_p, \quad \forall \omega \in \Omega \quad (41.1)$$

where $x_p \geq 0$ and $\forall p \in P$; that is, the sum of all the path flows between an origin/destination pair ω must be equal to the given demand d_ω .

In addition, the following conservation of flow equations must also hold:

$$f_a = \sum_{p \in P} x_p \delta_{ap}, \quad \forall a \in \mathcal{L} \quad (41.2)$$

where $\delta_{ap} = 1$, if link a is contained in path p , and 0, otherwise. Expression (41.2) states that the flow on link a is equal to the sum of all the path flows on paths p that contain (traverse) link a .

Equations (41.1) and (41.2) guarantee that the flows in the network (be they travelers, products, etc.) are conserved, that is, do not disappear (or are lost) in the network and arrive at the designated destinations from the origins.

Let c_a denote the user link cost associated with traversing link a , and let C_p denote the user cost associated with traversing the path p . Assume that the user link cost function is given by the *separable* function in which the cost on a link depends only on the flow on the link, that is,

$$c_a = c_a(f_a), \quad \forall a \in \mathcal{L} \quad (41.3)$$

where c_a is assumed to be continuous and an increasing function of the link flow f_a in order to model the effect of the link flow on the cost and, in particular, congestion.

The cost on a path is equal to the sum of the costs on the links that make up that path, that is,

$$C_p = \sum_{a \in \mathcal{L}} c_a(f_a) \delta_{ap}, \quad \forall p \in P \quad (41.4)$$

41.2.1.1 Transportation Network Equilibrium Conditions

In the case of the user-optimization (U-O) problem, one seeks to determine the path flow pattern x^* (and the corresponding link flow pattern f^*) which satisfies the conservation of flow Eqs. (41.1) and (41.2) and the nonnegativity assumption on the path flows and which also satisfies the transportation network equilibrium conditions given by the following statement. For each O/D pair $\omega \in \Omega$ and each path $p \in P_\omega$,

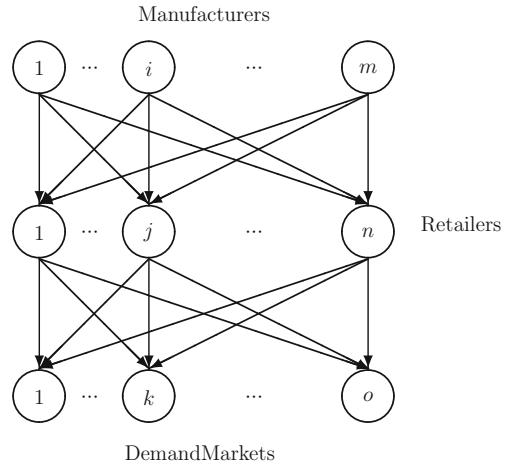
$$C_p \begin{cases} = \lambda_\omega, & \text{if } x_p^* > 0 \\ \geq \lambda_\omega, & \text{if } x_p^* = 0 \end{cases} \quad (41.5)$$

In the user-optimization problem, there is no explicit optimization criterion, since users of the transportation network system act independently, in a noncooperative manner, until they cannot improve on their situations unilaterally and, thus, an equilibrium is achieved, governed by the above equilibrium conditions. Conditions (41.5) are simply a restatement of Wardrop's (1952) first principle mathematically and mean that only those paths connecting an O/D pair will be used which have equal and minimal user costs. In Eq. (41.5) the minimal cost for O/D pair ω is denoted by λ_ω , and its value is obtained once the equilibrium flow pattern is determined. Otherwise, a user of the network could improve upon his situation by switching to a path with lower cost.

Beckmann et al. (1956) established that the solution to the network equilibrium problem, in the case of user link cost functions of the form Eq. (41.3), in which the cost on a link only depends on the flow on that link and is assumed to be continuous and an increasing function of the flow, could be obtained by solving the following optimization problem:

$$\text{Minimize} \quad \sum_{a \in \mathcal{L}} \int_0^{f_a} c_a(y) dy \quad (41.6)$$

Fig. 41.1 The multitered network structure of the supply chain



subject to

$$\sum_{p \in P_\omega} x_p = d_\omega, \quad \forall \omega \in \Omega \quad (41.7)$$

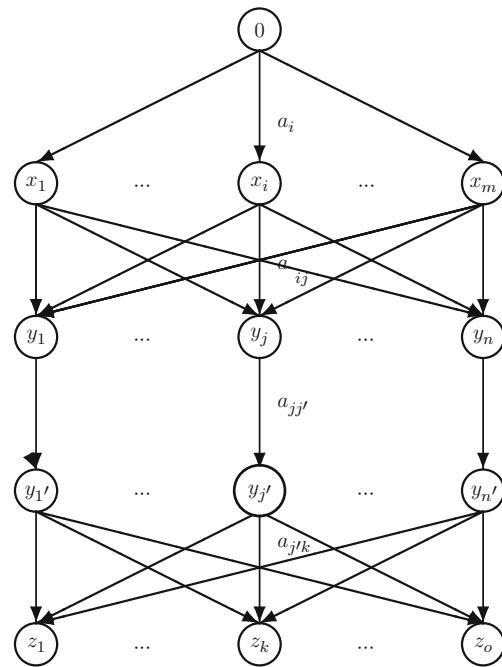
$$f_a = \sum_{p \in P} x_p \delta_{ap}, \quad \forall a \in \mathcal{L} \quad (41.8)$$

$$x_p \geq 0, \quad \forall p \in P \quad (41.9)$$

The objective function given by Eq. (41.6) is simply a device constructed to obtain a solution using general purpose convex programming algorithms. It does not possess the economic meaning of the objective function encountered in the system-optimization problem which will be recalled below. Note that in the case of separable, as well as nonseparable, but symmetric (which we come back to later) user link cost functions, the λ_ω term in Eq. (41.5) corresponds to the Lagrange multiplier associated with the constraint (41.7) for that O/D pair ω . However, in the case of nonseparable and asymmetric functions, there is no optimization reformulation of the transportation network equilibrium conditions (41.5), and the λ_ω term simply reflects the minimum user cost associated with the O/D pair ω at the equilibrium. As noted as early as Dafermos and Sparrow (1969), the above network equilibrium conditions also correspond to a Nash equilibrium (see Nash 1951). The equilibrium link flow pattern is unique for problem (41.6), subject to Eqs. (41.7)–(41.9), if the objective function (41.6) is strictly convex (for additional background on optimization theory.

It has also been established (cf. Nagurney (2006) and the references therein) that multitered supply chain network problems in which decision-makers (manufacturers, retailers, and even consumers) compete across a tier of the supply chain network but cooperate between tiers, as depicted in Fig. 41.1, could be transformed

Fig. 41.2 The supernetwork representation of supply chain network equilibrium



into a transportation network equilibrium problem using a supernetwork transformation, as in Fig. 41.2. In Fig. 41.2, the activities of manufacturing and retailer handling/storage are associated with the topmost and the third sets of links, respectively. The second and fourth sets of links from the top in Fig. 41.2 are the transportation links (as is the case with the links in Fig. 41.1). This connection provides us with a path flow efficiency interpretation of supply chain network equilibria. She utilized variational inequality theory (see below) to establish the equivalence.

41.2.2 The System-Optimized Problem

We now recall the system-optimized problem. As in the user-optimized problem of Section 41.2.1, the network $\mathcal{G} = [\mathcal{N}, \mathcal{L}]$, the demands associated with the origin/destination pairs, and the user link cost functions are assumed as given. In the system-optimized problem, there is a central controller who routes the flows in an optimal manner so as to minimize the total cost in the network. This problem has direct relevance to the management of operations of a supply chain.

The total cost on link a , denoted by $\hat{c}_a(f_a)$, is given by

$$\hat{c}_a(f_a) = c_a(f_a) \times f_a, \quad \forall a \in \mathcal{L} \quad (41.10)$$

that is, the total cost on a link is equal to the user link cost on the link times the flow on the link. As noted earlier, in the system-optimized problem, there exists a central

controller who seeks to minimize the total cost in the network system, which can correspond to a supply chain, where the total cost is expressed as

$$\sum_{a \in \mathcal{L}} \hat{c}_a(f_a) \quad (41.11)$$

and the total cost on a link is given by expression (41.10).

The system-optimization (S-O) problem is, thus, given by

$$\text{Minimize} \quad \sum_{a \in \mathcal{L}} \hat{c}_a(f_a) \quad (41.12)$$

subject to the same conservation of flow equations as for the user-optimized problem as well as the nonnegativity assumption of the path flows; that is, constraints (41.7), (41.8), and (41.9) must also be satisfied for the system-optimized problem.

The total cost on a path, denoted by \hat{C}_p , is the user cost on a path times the flow on a path, that is,

$$\hat{C}_p = C_p x_p, \quad \forall p \in P \quad (41.13)$$

where the user cost on a path, C_p , is given by the sum of the user costs on the links that comprise the path (as in Eq. (41.4)), that is,

$$C_p = \sum_{a \in \mathcal{L}} c_a(f_a) \delta_{ap}, \quad \forall a \in \mathcal{L} \quad (41.14)$$

In view of Eqs. (41.2), (41.3), and (41.4), one may express the cost on a path p as a function of the path flow variables, and, hence, an alternative version of the above system-optimization problem with objective function (41.12) can be stated in path flow variables only, where one has now the problem

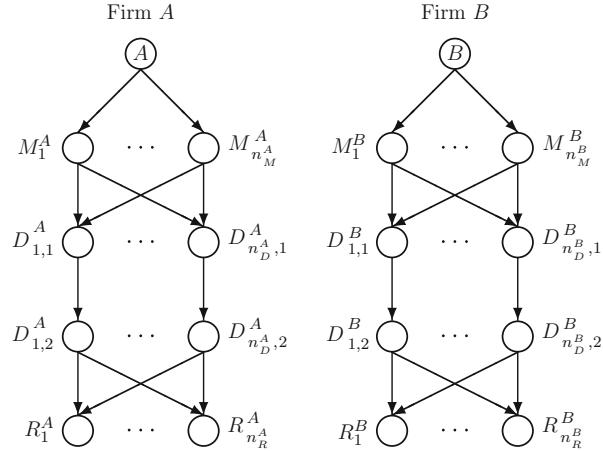
$$\text{Minimize} \quad \sum_{p \in P} C_p(x) x_p \quad (41.15)$$

subject to constraints (41.7) and (41.9).

41.2.2.1 System-Optimality Conditions

Under the assumption of increasing user link cost functions, the objective function (41.12) in the S-O problem is convex, and the feasible set consisting of the linear constraints (41.7)–(41.9) is also convex. Therefore, the optimality conditions, that is, the Kuhn-Tucker conditions, are as follows: for each O/D pair $\omega \in \Omega$ and each path $p \in P_\omega$, the flow pattern x (and corresponding link flow pattern f) satisfying Eqs. (41.7)–(41.9) must satisfy

Fig. 41.3 Case 0: firms A and B premerger



$$\hat{C}'_p \begin{cases} = \mu_\omega, & \text{if } x_p > 0 \\ \geq \mu_\omega, & \text{if } x_p = 0 \end{cases} \quad (41.16)$$

where \hat{C}'_p denotes the marginal of the total cost on path p , given by

$$\hat{C}_p = \sum_{a \in \mathcal{L}} \frac{\partial \hat{c}_a(f_a)}{\partial f_a} \delta_{ap} \quad (41.17)$$

evaluated in Eq. (41.16) at the solution and μ_ω is the Lagrange multiplier associated with constraint (41.7) for that O/D pair w .

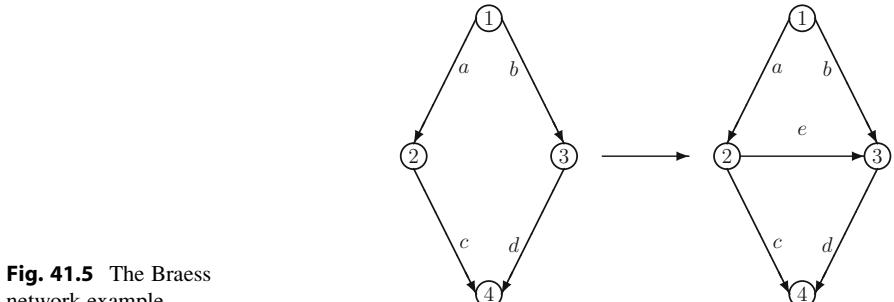
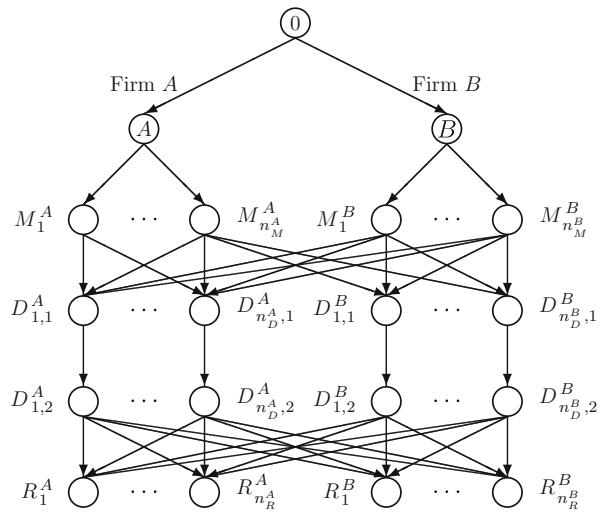
The system-optimization approach has been applied to supply chain networks in order to assess synergy associated with a possible merger or acquisition before such a decision, which may be very costly, is made. Nagurney and Qiang (2009) overview such an approach, which assesses the total cost prior to the merger and post.

The premerger supply chains corresponding to the individual firms, prior to the merger, are depicted in Fig. 41.3, whereas the post-merger supply chain network is given in Fig. 41.4. The topmost links correspond to the manufacturing links in Fig. 41.3, followed by the transportation links ending in the storage/distribution facility links and followed by additional shipment links to the demand markets. In Fig. 41.4, on the other hand, the topmost links represent the merger/acquisition with appropriate total cost functions assigned to those links.

41.2.3 The Braess Paradox

In order to illustrate the difference between user optimization and system optimization in a concrete example and to reinforce the above concepts, we now recall the well-known Braess (1968) paradox (see also Braess et al. 2005). Assume a network

Fig. 41.4 Post-merger network



as the first network depicted in Fig. 41.5 in which there are four nodes: 1, 2, 3, 4; four links: a, b, c, d ; and a single O/D pair $\omega_1 = (1, 4)$. There are, hence, two paths available to travelers between this O/D pair: $p_1 = (a, c)$ and $p_2 = (b, d)$.

The user link travel cost functions are

$$c_a(f_a) = 10f_a, \quad c_b(f_b) = f_b + 50, \quad c_c(f_c) = f_c + 50, \quad c_d(f_d) = 10f_d$$

Assume a fixed travel demand $d_{\omega_1} = 6$.

It is easy to verify that the equilibrium path flows are $x_{p_1}^* = 3$ and $x_{p_2}^* = 3$ and the equilibrium link flows are $f_a^* = 3$, $f_b^* = 3$, $f_c^* = 3$, $f_d^* = 3$, with associated equilibrium path travel costs: $C_{p_1} = c_a + c_c = 83$ and $C_{p_2} = c_b + c_d = 83$.

Assume now that, as depicted in Fig. 41.5, a new link “e,” joining node 2 to node 3, is added to the original network, with user link cost function $c_e(f_e) = f_e + 10$. The addition of this link creates a new path $p_3 = (a, e, d)$ that is available to the travelers. The travel demand d_{ω_1} remains at 6 units of flow. The original flow pattern $x_{p_1} = 3$ and $x_{p_2} = 3$ is no longer an equilibrium pattern, since, at this level of

flow, the user cost on path p_3 , $C_{p_3} = c_a + c_e + c_d = 70$. Hence, users on paths p_1 and p_2 would switch to path p_3 .

The equilibrium flow pattern on the new network is $x_{p_1}^* = 2$, $x_{p_2}^* = 2$, and $x_{p_3}^* = 2$, with equilibrium link flows $f_a^* = 4$, $f_b^* = 2$, $f_c^* = 2$, $f_e^* = 2$, and $f_d^* = 4$ and with associated equilibrium user path travel costs $C_{p_1} = 92$ and $C_{p_2} = 92$. Indeed, one can verify that any reallocation of the path flows would yield a higher travel cost on a path.

Note that the travel cost increased for every user of the network from 83 to 92 without a change in the travel demand!

The system-optimizing solution, on the other hand, for the first network in Fig. 41.5 is $x_{p_1} = x_{p_2} = 3$, with marginal total path costs given by $\hat{C}'_{p_1} = \hat{C}'_{p_2} = 116$. This would remain the system-optimizing solution, even after the addition of link e , since the marginal cost of path p_3 , \hat{C}'_{p_3} , at this feasible flow pattern is equal to 130.

The addition of a new link to a network cannot increase the total cost of the network system but can, of course, increase a user's cost since travelers act individually.

41.3 Models with Asymmetric Link Costs

In this section, we consider network models in which the user cost on a link is no longer dependent solely on the flow on that link. We present a fixed demand transportation network equilibrium model in Sect. 41.3.1 and an elastic demand one in Sect. 41.3.2.

We note that fixed demand supply chain network problems are relevant to applications in which there are good estimates of the demand as would be the case in certain healthcare applications. Elastic demand supply chain network problems can capture price sensitivity associated with the product and are used in profit-maximizing settings (cf. Nagurney 2006). Asymmetric link costs are relevant also in the case of competitive supply chain network equilibrium problems.

Assume that user link cost functions are now of a general form, that is, the cost on a link may depend not only on the flow on the link but on other link flows on the network, that is,

$$c_a = c_a(f), \quad \forall a \in \mathcal{L} \quad (41.18)$$

In the case where the symmetry assumption exists, that is, $\frac{\partial c_a(f)}{\partial f_b} = \frac{\partial c_b(f)}{\partial f_a}$, for all links $a, b \in \mathcal{L}$, one can still reformulate the solution to the network equilibrium problem satisfying equilibrium conditions (41.5) as the solution to an optimization problem, albeit, again, with an objective function that is artificial and simply a mathematical device. However, when the symmetry assumption is no longer satisfied, such an optimization reformulation no longer exists, and one must appeal to *variational inequality theory* (cf. Nagurney (1999) and the references therein).

Models of supply chains and transportation networks with asymmetric cost functions are important since they allow for the formulation, qualitative analysis, and, ultimately, solution to problems in which the cost on a link may depend on the flow on another link in a different way than the cost on the other link depends on that link's flow.

It was in the domain of such network equilibrium problems that the theory of finite-dimensional variational inequalities realized its earliest success, beginning with the contributions of Smith (1979) and Dafermos (1980). For an introduction to the subject, as well as applications ranging from transportation network and spatial price equilibrium problems to financial equilibrium problems, see the book by Nagurney (1999). Below we present variational inequality formulations of both fixed demand and elastic demand network equilibrium problems.

The system-optimization problem, in turn, in the case of nonseparable (cf. Eq. (41.18)) user link cost functions becomes (see also Eq. (41.12))

$$\text{Minimize} \quad \sum_{a \in \mathcal{L}} \hat{c}_a(f) \quad (41.19)$$

subject to Eqs. (41.7)–(41.9), where $\hat{c}_a(f) = c_a(f) \times f_a$ and $\forall a \in \mathcal{L}$.

The system-optimality conditions remain as in Eq. (41.16), but now the marginal of the total cost on a path becomes, in this more general case,

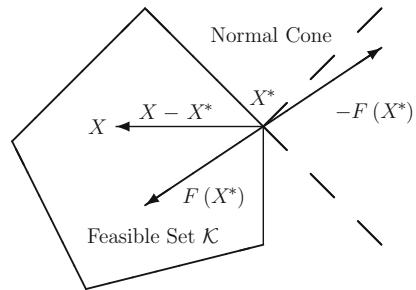
$$\hat{C}'_p = \sum_{a,b \in \mathcal{L}} \frac{\partial \hat{c}_b(f)}{\partial f_a} \delta_{ap}, \quad \forall p \in P \quad (41.20)$$

41.3.1 Variational Inequality Formulations of Fixed Demand Problems

As mentioned earlier, in the case where the user link cost functions are no longer symmetric, one cannot compute the solution to the U-O, that is, to the network equilibrium, problem using standard optimization algorithms. We emphasize, again, that such general cost functions are very important from an application standpoint since they allow for asymmetric interactions on the network. For example, allowing for asymmetric cost functions permits one to handle the situation when the flow on a particular link affects the cost on another link in a different way than the cost on the particular link is affected by the flow on the other link.

First, the definition of a variational inequality problem is recalled. For further background, theoretical formulations, derivations, and the proofs of the results below, see the books by Nagurney (1999) and by Nagurney and Dong (2002) and the references therein. We provide the variational inequality of the network equilibrium conditions in path flows as well as in link flows since different formulations suggest different computational methods for solution.

Fig. 41.6 Geometric interpretation of VI (F, \mathcal{K})



Specifically, the variational inequality problem (finite-dimensional) is defined as follows:

Definition 1: Variational Inequality Problem

The finite-dimensional variational inequality problem, VI (F, \mathcal{K}) , is to determine a vector $X^* \in \mathcal{K}$ such that

$$\langle F(X^*)^T, X - X^* \rangle \geq 0, \quad \forall X \in \mathcal{K} \quad (41.21)$$

where F is a given continuous function from \mathcal{K} to R^N , \mathcal{K} is a given closed convex set, and $\langle \cdot, \cdot \rangle$ denotes the inner product in R^N .

Variational inequality Eq. (41.21) is referred to as being in *standard form*. Hence, for a given problem, typically an *equilibrium* problem, one must determine the function F that enters the variational inequality problem, the vector of variables X , as well as the feasible set \mathcal{K} .

The variational inequality problem contains, as special cases, such well-known problems as systems of equations, optimization problems, and complementarity problems. Thus, it is a powerful unifying methodology for equilibrium analysis and computation and continues to be utilized for the formulation, analysis, and solution of a spectrum of supply chain network problems (cf. Nagurney 2006).

A geometric interpretation of the variational inequality problem VI (F, \mathcal{K}) is given in Fig. 41.6. Specifically, $F(X^*)$ is “orthogonal” to the feasible set \mathcal{K} at the point X^* .

Theorem 1: Variational Inequality Formulation of Network Equilibrium with Fixed Demands: Path Flow Version

A vector $x^* \in K^1$ is a network equilibrium path flow pattern, that is, it satisfies equilibrium conditions (41.5) if and only if it satisfies the variational inequality problem:

$$\sum_{\omega \in \Omega} \sum_{p \in P_\omega} C_p(x^*) \times (x - x^*) \geq 0, \quad \forall x \in K^1 \quad (41.22)$$

or, in vector form,

$$\langle C(x^*)^T, x - x^* \rangle \geq 0, \quad \forall x \in K^1 \quad (41.23)$$

where C is the n_P -dimensional vector of path user costs and K^1 is defined as $K^1 \equiv \{x \geq 0, \text{ such that Eq. (41.7) holds}\}$.

Theorem 2: Variational Inequality Formulation of Network Equilibrium with Fixed Demands: Link Flow Version

A vector $f^* \in K^2$ is a network equilibrium link flow pattern if and only if it satisfies the variational inequality problem:

$$\sum_{a \in \mathcal{L}} c_a(f^*) \times (f_a - f_a^*) \geq 0, \quad \forall f \in K^2 \quad (41.24)$$

or, in vector form,

$$\langle c(f^*)^T, f - f^* \rangle \geq 0, \quad \forall f \in K^2 \quad (41.25)$$

where c is the n_L -dimensional vector of link user costs and K^2 is defined as $K^2 \equiv \{f \mid \text{there exists an } x \geq 0 \text{ and satisfying Eqs. (41.7) and (41.8)}\}$.

One may put variational inequality Eq. (41.23) into standard form Eq. (41.21) by letting $F \equiv C$, $X \equiv x$, and $K \equiv K^1$. One may also put variational inequality Eq. (41.25) into standard form where now $F \equiv c$, $X \equiv f$, and $K \equiv K^2$. Hence, fixed demand transportation network equilibrium problems in the case of asymmetric user link cost functions can be solved as variational inequality problems, as given above.

The theory of variational inequalities (see Kinderlehrer and Stampacchia 1980; Nagurney 1999) allows one to qualitatively analyze the equilibrium patterns in terms of existence, uniqueness, as well as sensitivity and stability of solutions and to apply rigorous algorithms for the numerical computation of the equilibrium patterns. Variational inequality algorithms usually resolve the variational inequality problem into series of simpler subproblems, which, in turn, are often optimization problems, which can then be effectively solved using a variety of algorithms.

We emphasize that the above network equilibrium framework is sufficiently general to also formalize the entire transportation planning process (consisting

of origin selection, or destination selection, or both, in addition to route selection, in an optimal fashion) as path choices over an appropriately constructed *abstract* network or supernetwork. Further discussion can be found in the books by Nagurney (1999, 2000) and Nagurney and Dong (2002) who also developed more general models in which the costs (as described above) need not be separable nor asymmetric.

41.3.2 Variational Inequality Formulations of Elastic Demand Problems

We now describe a general network equilibrium model with elastic demands due to Dafermos (1982), but we present the single-modal version for simplicity. It is assumed that one has associated with each O/D pair ω in the network a travel disutility function λ_ω , where here the general case is considered in which the disutility may depend upon the entire vector of demands, which are no longer fixed, but are now variables, that is,

$$\lambda_\omega = \lambda_\omega(d), \quad \forall \omega \in \Omega \quad (41.26)$$

where d is the J -dimensional vector of the demands.

The notation is as described earlier, except that here we also consider user link cost functions which are general, that is, of the form Eq. (41.18). The conservation of flow equations (see also Eqs. (41.1) and (41.2)), in turn, is given by

$$f_a = \sum_{p \in P} x_p \delta_{ap}, \quad \forall a \in \mathcal{L} \quad (41.27)$$

$$d_\omega = \sum_{p \in P_\omega} x_p, \quad \forall \omega \in \Omega \quad (41.28)$$

$$x_p \geq 0, \quad \forall p \in P \quad (41.29)$$

In the elastic demand case, the demands in expression (41.28) are variables and no longer given, in contrast to the fixed demand expression in Eq. (41.1).

The network equilibrium conditions (see also Eq. (41.5)) take on in the elastic demand case the following form. For every O/D pair $\omega \in \Omega$ and each path $p \in P_\omega$, a vector of path flows and demands (x^*, d^*) satisfying Eqs. (41.28) and (41.29) (which induces a link flow pattern f^* through Eq. (41.27)) is a network equilibrium pattern if it satisfies

$$C_p(x^*) \begin{cases} = \lambda_\omega(d^*), & \text{if } x_p^* > 0 \\ \geq \lambda_\omega(d^*), & \text{if } x_p^* = 0 \end{cases} \quad (41.30)$$

Equilibrium conditions (41.30) state that the costs on used paths for each O/D pair are equal and minimal and equal to the disutility associated with that O/D pair. Costs on unutilized paths can exceed the disutility. Observe that in the elastic demand model users of the network can forego travel altogether for a given O/D pair if the user costs on the connecting paths exceed the travel disutility associated with that O/D pair. This model, hence, allows one to ascertain the attractiveness of different O/D pairs based on the ultimate equilibrium demand associated with the O/D pairs. In addition, this model can handle such situations as the equilibrium determination of employment location and route selection, or residential location and route selection, or residential and employment selection as well as route selection through the appropriate transformations via the addition of links and nodes and given, respectively, functions associated with the residential locations, the employment locations, and the network overall (cf. Nagurney 1999; Nagurney and Dong 2002).

In the next two theorems, both the path flow version and the link flow version of the variational inequality formulations of the network equilibrium conditions (41.30) are presented. These are analogues of the formulations Eqs. (41.22) and (41.23) and (41.24) and (41.25), respectively, for the fixed demand model and are due to Dafermos (1982).

Theorem 3: Variational Inequality Formulation of Network Equilibrium with Elastic Demands: Path Flow Version

A vector $(x^*, d^*) \in K^3$ is a network equilibrium path flow pattern, that is, it satisfies equilibrium conditions (41.30) if and only if it satisfies the variational inequality problem:

$$\sum_{\omega \in \Omega} \sum_{p \in P_\omega} C_p(x^*) \times (x - x^*) - \sum_{\omega \in \Omega} \lambda_\omega(d^*) \times (d_\omega - d_\omega^*) \geq 0 \quad (41.31)$$

$$\forall (x, d) \in K^3,$$

or, in vector form,

$$\left\langle C(x^*)^T, x - x^* \right\rangle - \left\langle \lambda(d^*)^T, d - d^* \right\rangle \geq 0, \quad \forall (x, d) \in K^3 \quad (41.32)$$

where λ is the J -dimensional vector of disutilities and K^3 is defined as $K^3 \equiv \{x \geq 0, \text{ such that Eq. (41.28) holds}\}$.

Theorem 4: Variational Inequality Formulation of Network Equilibrium with Elastic Demands: Link Flow Version

A vector $(f^*, d^*) \in K^4$ is a network equilibrium link flow pattern if and only if it satisfies the variational inequality problem:

$$\sum_{a \in \mathcal{L}} c_a(f^*) \times (f_a - f_a^*) - \sum_{\omega \in \Omega} \lambda_\omega(d^*) \times (d_\omega - d_\omega^*) \geq 0 \quad (41.33)$$

$$\forall (f, d) \in K^4,$$

or, in vector form,

$$\left\langle c(f^*)^T, f - f^* \right\rangle - \left\langle \lambda(d^*)^T, d - d^* \right\rangle \geq 0, \quad \forall (f, d) \in K^4 \quad (41.34)$$

where $K^4 \equiv \{(f, d), \text{ such that there exists an } x \geq 0 \text{ satisfying Eqs. (41.27), (41.28)}\}$.

Under the symmetry assumption on the disutility functions, that is, if $\frac{\partial \lambda_w}{\partial d_\omega} = \frac{\partial \lambda_\omega}{\partial d_w}$, for all w, ω , in addition to such an assumption on the user link cost functions (see following Eq. (41.18)), one can obtain (see Beckmann et al. 1956) an optimization reformulation of the network equilibrium conditions (41.30), which in the case of separable user link cost functions and disutility functions is given by

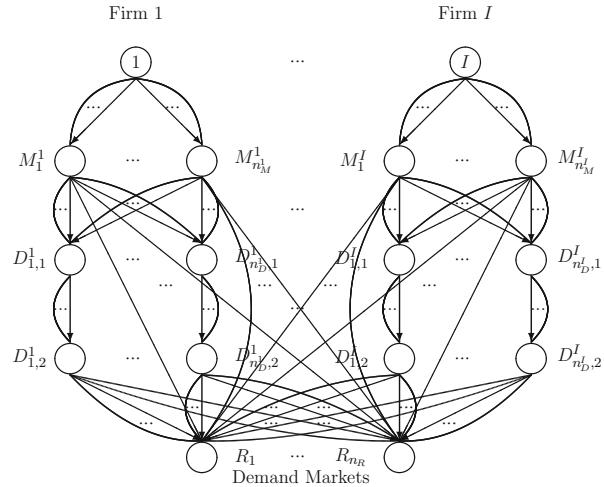
$$\text{Minimize} \quad \sum_{a \in \mathcal{L}} \int_0^{f_a} c_a(y) dy - \sum_{\omega \in \Omega} \int_0^{d_\omega} \lambda_\omega(z) dz \quad (41.35)$$

subject to Eqs. (41.27)–(41.29).

Variational inequality theory has become a fundamental methodological framework for the formulation and solution of competitive supply chain problems in which the governing concept is that of Nash equilibrium (see, e.g., Masoumi et al. 2012).

In Fig. 41.7, a competitive supply chain network is depicted in which the firms have vertically integrated supply chains but compete in common demand markets. The topmost links represent manufacturing activities at different plants with different such links denoting alternative manufacturing technologies. The second set of links from the top reflects transportation, and alternative links depict the possibility of alternative modes of transportation. The next set of links corresponds to storage at the distribution centers and the final set of links the transportation to the demand markets. Here we also use multiple links to denote alternative technologies

Fig. 41.7 The competitive supply chain network topology



and transportation modes, respectively. The costs on the links can be separable or not and asymmetric, depending on the specific product application. Product differentiation and branding has also been incorporated into such supply chain networks using variational inequality theory. Observe that in the supply chain network depicted in Fig. 41.7, direct shipments from the manufacturing plants to the demand points/retailers are allowed and depicted by the corresponding links.

Finally, it is important to emphasize that the dynamics of the underlying interactions can be formulated and has been done so using projected dynamical systems (Nagurney and Zhang 1996).

41.4 Conclusions

In this chapter, we have highlighted some of the major advances in supply chains and transportation networks, with a focus on the common elements as to the theoretical frameworks and underlying behavioral principles. We have also argued that the foundations of supply chains as network systems can be found in the regional science and spatial economics literature.

Specifically, we have discussed how the concepts of system optimization and user optimization have underpinned transportation network models and, more recently, have evolved to enable the formulation of supply chain network problems operating (and managed) under centralized or decentralized, that is, competitive, decision-making behavior.

We have also highlighted some of the principal methodologies, including variational inequality theory, that have enabled the development not only of advanced transportation network equilibrium models but also supply chain network equilibrium models.

We have aimed to include both primary references as well as tertiary references; the interested reader can delve further, at his/her convenience and according to interest.

In conclusion, transportation network concepts, models, and accompanying methodologies have enabled the advancement of supply chain network models from a system-wide and holistic perspective.

References

- Beckmann MJ, McGuire CB, Winsten CB (1956) Studies in the economics of transportation. Yale University Press, New Haven
- Boone T, Jayaraman V, Ganeshan R (2012) Sustainable supply chains: models, methods, and public policy implications. Springer, New York
- Boyce DE, Mahmassani HS, Nagurney A (2005) A retrospective on Beckmann, McGuire, and Winsten's studies in the economics of transportation. *Pap Reg Sci* 84:85–103
- Braess D (1968) Über ein paradoxon der verkehrsplanung. *Unternehmensforschung* 12:258–268
- Braess D, Nagurney A, Wakolbinger T (2005) On a paradox of traffic planning, translation of the original D. Braess paper from German to English. *Transp Sci* 39:446–450
- Dafermos S (1980) Traffic equilibrium and variational inequalities. *Transp Sci* 14:42–54
- Dafermos S (1982) The general multimodal network equilibrium problem with elastic demand. *Networks* 12:57–72
- Dafermos SC, Sparrow FT (1969) The traffic assignment problem for a general network. *J Res Nat Bur Stand* 73B:91–118
- Handfield RB, Nichols EL Jr (1999) Introduction to supply chain management. Prentice-Hall, Englewood Cliffs
- Isard W (1954) Location theory and trade theory: short-run analysis. *Q J Econ* 68:305–320
- Kinderlehrer D, Stampacchia G (1980) An introduction to variational inequalities and their applications. Academic Press, New York
- Masoumi AH, Yu M, Nagurney A (2012) A supply chain generalized network oligopoly model for pharmaceuticals under brand differentiation and perishability. *Transp Res E* 48:762–780
- Nagurney A (1999) Network economics: a variational inequality approach, second and revised edition. Kluwer, Dordrecht
- Nagurney A (2000) Sustainable transportation networks. Edward Elgar, Cheltenham
- Nagurney A (2006) Supply chain network economics: dynamics of prices, flows and profits. Edward Elgar, Cheltenham
- Nagurney A (2007) Mathematical models of transportation and networks. In: Zhang W-B (ed) Encyclopedia of life support systems (EOLSS), Mathematical models in economics. United Nations Educational, Scientific and Cultural Organization (UNESCO), Paris
- Nagurney A (2010) Optimal supply chain network design and redesign at minimal total cost and with demand satisfaction. *Int J Prod Econ* 128:200–208
- Nagurney A, Dong J (2002) Supernetworks: decision-making for the Information Age. Edward Elgar, Cheltenham
- Nagurney A, Dong J, Zhang D (2002) A supply chain network equilibrium model. *Transp Res E* 38:281–303
- Nagurney A, Qiang Q (2009) Fragile networks: identifying vulnerabilities and synergies in an uncertain world. Wiley, Hoboken
- Nagurney A, Zhang D (1996) Projected dynamical systems and variational inequalities with applications. Kluwer, Norwell
- Nash JF (1951) Noncooperative games. *Ann Math* 54:286–298
- Ohlin B (1933) Interregional and international trade. Harvard University Press, Cambridge, MA
- Patriksson M (1994) The traffic assignment problem. VSP, Utrecht

- Ran B, Boyce DE (1996) Modeling dynamic transportation networks, 2 revisedth edn. Springer, Berlin
- Samuelson PA (1952) Spatial price equilibrium and linear programming. *Am Econ Rev* 42:283–303
- Sheffi Y (1985) Urban transportation networks. Prentice-Hall, Englewood Cliffs
- Smith MJ (1979) Existence, uniqueness, and stability of traffic equilibria. *Transp Res B* 13:259–304
- Takayama T, Judge GG (1971) Spatial and temporal price and allocation models. North-Holland, Amsterdam
- Wardrop JG (1952) Some theoretical aspects of road traffic research. *Proc Inst Civil Eng* 1(II):325–378
- Zhang D, Dong J, Nagurney A (2003) A supply chain network economy: modeling and qualitative analysis. In: Nagurney A (ed) Innovations in financial and economic networks. Edward Elgar, Cheltenham, pp 197–213

Aura Reggiani

Contents

42.1	Introduction	812
42.2	Complexity and Spatial Networks	812
42.3	Static Complexity and Models	815
42.3.1	Preface	815
42.3.2	Static Complexity and Static Models in Spatial Economic Analysis	815
42.4	Dynamic Complexity and Models	818
42.4.1	Simple Models vs. Dynamic Complexity	818
42.4.2	Less Simple Models vs. Dynamic Complexity	821
42.4.3	Concluding Remarks	822
42.5	Complexity and Network Analysis	823
42.5.1	Preface	823
42.5.2	Simple Network Models: Random and Scale-Free Networks	824
42.5.3	Concluding Remarks	826
42.6	Spatial Economics and Network Analysis: Connectivity, Emergence, and Resilience	827
42.7	Conclusions	829
	References	831

Abstract

The modern spatial economy has a global “networked” character that is generating important socioeconomic and political changes. In this respect, new forms of connectivity play a significant role through their dynamic and complex interplay with the economic and political driving forces behind globalization. In analyzing such impacts, it is useful to consider the tools and models that have been adopted in regional economics as well as in other disciplines. In this

A. Reggiani

Department of Economics, University of Bologna, Bologna, Italy

e-mail: aura.reggiani@unibo.it

context, it is also necessary to reflect on complexity theory and on the models able to map out the complex interconnected spatial networks.

This chapter begins with a concise review of the most important definitions of complexity, in the light of their relations with spatial networks. There follows an exploration of the main findings from two “close” disciplines, that is, spatial economics and network science, with reference to their associated approaches and modeling tools which are able to grasp complexity from, respectively, the behavioral and the network structure viewpoint. The emerging discussion – with reference to both static and dynamic frameworks through the lens of complexity issues – indicates that (i) a formal correspondence between the fundamental spatial economic models and network models exists and (ii) this correspondence highlights the “simplicity” of the laws underlying complex spatial networks.

42.1 Introduction

The structure and the development of spatial networks, that is, networks where space – in the form of distance friction and/or transportation/communication costs/utilities – assumes a fundamental role in the economic activities, are currently experiencing unpredictable changes and following diverse paths. These uncertainties are mainly the result of the increasing connectivity, at all scale levels, of information and communication – and in general of economic – systems in our society.

In this complex and heterogeneous landscape, a central issue of research is the adoption and validation of approaches and methodologies able to grasp these aspects of economic uncertainty and discontinuity and overcome the current difficulties of carrying out reliable forecasts. In this vein, concepts, such as dynamics, complexity, connectivity, emergence and self-organization, vulnerability, and resilience – which have received a great deal of attention in recent decades – have been shown to provide scientists with a powerful framework for viewing the complex spatial economic transformation processes.

In this chapter, we discuss some of these issues, by focusing on the main modeling tools which have been adopted in the scientific literature in order to investigate the complex dynamics of this networked space economy. This concise chapter will be based on the exploration of the main findings from two “close” fields: spatial economics and network science. First, Sect. 42.2 outlines the essential points concerning the definition of complexity, in order to provide a historical and conceptual background to the subsequent analyses. The fundamental models in spatial economic analysis, that is, the main static and dynamic models able to grasp complexity from the behavioral viewpoint, are then presented in Sects. 42.3 and 42.4, in order to identify the similarities and synergies with the tools employed in network analysis (from the network structure viewpoint), in the light of their simplicity. The network models are then examined in Sect. 42.5, while this dual analysis (spatial economics vs. network analysis) will be dealt with briefly in Sect. 42.6, with reference to the central concepts of complexity, namely, connectivity, emergence, and resilience. Finally, Sect. 42.7 concludes the chapter with

suggestions for new paths in future research: mainly the necessity for a blend of advanced theories and approaches belonging to regional economics and complex network theory, with the final aim of assessing such models for possible use in an operational setting.

42.2 Complexity and Spatial Networks

“Complexity has turned out to be very difficult to define. The dozens of definitions that have been offered all fall short in one respect or another, classifying something as complex which we intuitively would see as simple, or denying an obviously complex phenomenon the label of complexity” (Heylighen 1996, p. 1). As Heylighen argued, defining complexity is fraught with difficulties. Horgan, in his 1996 article entitled “From Complexity to Perplexity,” mentions 31 definitions of complexity and associated concepts (Reggiani 2004). Given this wide arena and production of works on the meaning of complexity, it is worth examining the etymology of the term “complexity.” As Heylighen (1996) has noted, the original Latin word *complexus* means “entwined,” “twisted together”; furthermore, the Oxford English Dictionary defines something as “complex” if it is “made of (usually several) closely connected parts.” From these definitions it is clear that the term “complexity” embeds both the assemblage of different units in a system and their intertwined dynamics. In other words, the term “complexity” is strictly related to the concept of networks.

Several definitions also exist concerning the term “network.” Let us then also consider the etymology of the term “network.” Literally, the notion of network refers to “operations via nets.” In this context, Nijkamp and Reggiani (1998) argue that spatial networks may be interpreted as an ordered connectivity structure for spatial communication and transportation which is characterized by the existence of main nodes which act as receivers or senders (push and pull centers) and which are connected by means of corridors and edges. The relevance of the dynamic function of the (spatial) network via organized linkage patterns is embedded in this definition. Here, it is interesting to recall the simple definition by Barthélémy (2010, p. 3): “Loosely speaking, spatial networks are networks for which the nodes are located in a space equipped with a metric.”

The relationship between complexity and (spatial) networks can be structured in the following way, on the basis of Casti’s (1979) classification of complexity:

- *Static complexity*: refers to the network configuration, where the components are put together in an interrelated and intricate way. Network configuration concerns, for example, the number and type of hierarchical structures, the type of the connectivity patterns, the variety of components, and the strength of the interactions. Clearly, static complexity can be roughly measured by the above-mentioned variables.
- *Dynamic complexity*: concerns the dynamic network behavior governed by nonlinearities in the interacting components. Here, two rough measures can be the computational complexity and the evolutionary complexity. The latter

measure can be carried out by means of appropriate nonlinear models, like chaos models, in particular, and evolutionary models in general, which are able to map out the dynamic (random) network patterns.

If we consider the synthetic, but exhaustive, definition of a complex system formulated by Simon (1962, p. 468) as a “large number of parts that interact in a nonsimple way” vs. Casti’s (1979, p. 97) definition of complexity: “The primary idea of complexity concerns the mapping of a system’s non-intuitive behaviour, particularly the evolutionary patterns of connections among interacting components of a system whose long-run behaviour is hard to predict,” we can extract a further important element which – in addition to the network concept – characterizes complexity, namely, the (random) dynamic behavior, which is difficult to predict.

It should be noted that a previous interesting classification of complexity was provided by Weaver in 1948, in his article “Science and Complexity,” as follows:

- *Disorganized complexity*: concerns a situation in which the number of interacting variables is very large and in which each of the many variables has a behavior which is individually erratic or perhaps totally unknown. For this type of problem, the statistical methods hold the key. In this context, Weaver provides the examples of a large telephone exchange, the financial stability of a life insurance company, and the motion of the atoms and stars, which suggest a “whole array of practical applications and statistical techniques based on disorganized complexity” (Weaver 1948, p. 538).
- *Organized complexity*: concerns a situation in which the number of variables is moderate and their interrelationships cannot fully be captured in probabilistic statistics. Weaver considers here the “middle” regions, where the number of variables is moderate: large compared to two but small compared to the number of atoms. For example, the reproduction mechanism or the chemical reactions, as well as some macroeconomic relationships (e.g., on which variable the price of wheat depends), are problems “which involve dealing simultaneously with a sizable number of factors which are interrelated into an organic whole” (Weaver 1948, p. 539).

The concepts of networks and erratic behavior are encapsulated in Weaver’s two definitions, where, in addition, the concept of simplicity can be identified as follows: (a) statistical methods are a way of “decoding” the disorganized complexity and (b) the “organic wholes,” with their parts in close interrelation, represent another type of simplicity approach. In other words, already in Weaver, as also later on in Casti, the simplicity concept appears to be intrinsically related to the concept of complexity, since it seems the only way of “governing” complexity from the scientific viewpoint. The issue of harnessing complexity has also been tackled by Axelrod and Cohen (2000), with reference to the difficult task of making predictions in complex settings and thus to the necessity of providing a device for channeling the complexity of a system into desirable change.

Even though the above classifications can help in the discussion on how to define and tackle complexity, the objective of identifying a unified theory of complexity is still open. On the one hand, both systems and network theory may help in defining analytical, and hence measurable, complexity, although it remains difficult to

capture inherent behavioral complexity. On the other hand, both systems and economic theory may help in trying to understand the dynamic complexity of spatial economic phenomena, by analyzing appropriate dynamic models. The idea of dynamic systems with complex landscapes has been advocated by Krugman (1994), as a unifying theme in a number of research fields in the last decades. In other words, an interdisciplinary approach, able to fill the gap between socioeconomic and physical science, might grasp the common universal principles which can create a kind of unified science of complexity. A first step in this respect is to look at complexity in terms of two interrelated approaches: spatial economic analysis (Sects. 42.3 and 42.4) and network analysis (Sect. 42.5).

42.3 Static Complexity and Models

42.3.1 Preface

Spatial economics seeks to identify the factors governing the distribution/location of economic activity over space; thus, the space economy can be interpreted as a well-functioning economic system enriched with the element of space. Here, the complex evolution of interrelated spatial economic networks (e.g., transport and communication networks, industrial and financial networks, socioeconomic and organizational networks) plays a crucial role in the economic growth of regions/countries and in the related forecasting analyses. The contrast between slow (e.g., the evolution of physical infrastructure networks) and fast dynamics (e.g., the evolution of digital communication networks) points out the unpredictable (dynamic) character of such spatial interconnected networks and calls for a complexity approach able to understand the underlying forces and emerging processes.

In this landscape of multifaceted developments of spatial economic systems, a fascinating scientific question is whether the models which have been formed on the basis of the spatial economic analysis are still useful and in what respect, with reference to these new concepts of complexity and networks. A concise review of the fundamental models in spatial economic analysis, with reference to static complexity, is provided next in Sect. 42.3. The related dynamic framework through the lens of complexity issues will be discussed in Sect. 42.4.

42.3.2 Static Complexity and Static Models in Spatial Economic Analysis

By considering a *static framework*, some fundamental models in spatial economics can be summarized in their historical evolution as follows:

- (a) The rank-size rule/Zipf's law (Zipf 1949)
- (b) Gravity models (Isard 1956)
- (c) Spatial interaction models (Wilson 1970)
- (d) Discrete choice models (McFadden 1974)

These four types of models are all static models, with very simple formulations. Their common characteristic is to be *spatial models* (where space is (generally) represented in the form of a system of discrete locations (zones) at a certain level of resolution), at the aggregate level for models (a), (b), and (c) and at the disaggregate level for model (d). It is well known that we can observe an *analytical compatibility* between the spatial interaction model (SIM) and all the other models (a), (b), and (d) (Batty 2010; Reggiani 2004; Reggiani and Nijkamp 2009).

The general form of a (doubly-constrained) SIM reads as follows:

$$F_{ij} = A_i B_j O_i D_j f(\beta, c_{ij}), \quad i = 1, \dots, I \quad j = 1, \dots, J \quad (42.1)$$

where F_{ij} represents the total number of flows (physical or virtual) between the origin i and the destination j ; O_i and D_j are the stock variables (e.g., population size, workplaces) in the places of origin and destination; c_{ij} are the generalized interaction costs; the term $f(\beta, c_{ij})$ is the impedance function, measuring separation effects between i and j ; and β is the cost-sensitivity parameter. It should be noted that different types of impedance functions can be used, according to the type of spatial structure under analysis (e.g., the negative exponential for homogeneous centers/nodes in the spatial network, the negative power in the presence of large agglomerations/metropolitan areas, etc.: for a review, see Fotheringham and O'Kelly 1989; Reggiani and Nijkamp 2009). The terms A_i and B_j are balancing factors, equal to:

$$A_i = 1 / \sum_j B_j D_j f(\beta, c_{ij}), \quad B_j = 1 / \sum_i A_i O_i f(\beta, c_{ij}) \quad (42.2)$$

both derived from the respective additivity conditions:

$$\sum_j F_{ij} = O_i, \quad \sum_i F_{ij} = D_j \quad (42.3)$$

Model (42.1) can be derived as a probabilistic approach based on statistical equilibrium concepts (Wilson 1970). Wilson, in fact, demonstrated that SIM (42.1) can be derived from a mathematical optimization problem, by maximizing an entropy function, and can thus be seen as an optimum systems solution. The SIM (42.1) can be then perceived as the equilibrium state solution in the network of erratic movements. This approach provided a macrobehavioral context to SIMs, given that entropy can be interpreted in terms of a generalized cost function for spatial interaction behavior (Nijkamp and Reggiani 1992).

Consequently, the SIM (42.1) appears to be a suitable instrument/model able to deal with static complexity (e.g., in the presence of a high number of (origin/destination) cells in the network), that is, when the dynamic complex network is in equilibrium.

Interestingly, the SIM (42.1) appears to be the focal “model” among the others, since it emerges from different theoretical roots. In particular, the SIM is linked to:

- The gravity principle (Newton's law) (Isard 1956) and to the rank-size rule (Reggiani and Nijkamp 2009)

- Statistical information principles and entropy maximization (Wilson 1970)
- The logit model and thus to microeconomic theory (stochastic utility maximization) (Anas 1983).

Concerning the first methodological link, that is, the SIM and gravity model/rank-size rule, let us consider the unconstrained version of the SIM:

$$F_{ij} = KO_i D_j f(\beta, c_{ij}) \quad (42.4)$$

where K is a scaling factor, which has to be calibrated.

Model (42.4) clearly reflects Newton's gravity law, in the network of the cost/distance relations $f(\beta, c_{ij})$ between the masses (in our case populations O_i and workplaces D_j) (see, among others, Sen and Smith 1995).

Let us then examine the rank-size formulation, which reads as follows:

$$P_j = GR_j^{-q} \quad (j = 1, \dots, J) \quad (42.5)$$

where P_j is a given size of a city population j , R_j is the related rank of the city j , q is the elasticity parameter, and G is a positive constant (usually the population of the biggest city). In the particular case of $q = 1$, the rank-size rule boils down to the well-known Zipf's law (Zipf 1949) and thus to a perfect equilateral hyperbola where the agglomeration/centripetal forces (influencing the masses or population P_j in the network) are in equilibrium with the opposing diversification/centrifugal forces (influencing the rank R_j).

Furthermore, the SIM (42.4) can be formulated as follows:

$$F_{ij}/O_i D_j = Kf(\beta, c_{ij}) \quad (42.6)$$

by showing its link to the rank-size rule (42.5) (see also Batty 2010). The rank-size rule (42.5), in fact, appears to be compatible with the spatial/gravity interaction (Eq. (42.6)), since, like the SIM, it can be derived from an entropy maximization approach, in particular by maximizing the most likely combination of population stocks P_j from among a very large number of realizations of independent microllevel outcomes. In this interpretation, entropy essentially refers to the maximum probability of decentralization among random population centers, and the rank R_j can assume an economic value, being interpreted as the shadow cost. In other words, entropy can be considered as an indicator of the static network complexity, whereas the SIM (42.1) and the rank-size rule (42.5) represent the optimal equilibrium solutions. It is interesting to note that Zipf had already provided a form of "cost/distance" interpretation of the rank R_j , by arguing: "In other words, and in general, the most frequently used good services tend to be the cheapest and the nearest, and the reverse" (Zipf 1949 p. 371).

In addition, the rank-size rule (42.5) can be written as:

$$G = P_j R_j^q \quad (j = 1, \dots, J) \quad (42.7)$$

Surprisingly, Eq. (42.7) resembles Einstein's law (1905):

$$E = mc^2 \quad (42.8)$$

In Einstein's expression (42.8), E is energy, m is mass, and c is the speed of light, which – like the rank R_j – may clearly have an economic value; here the coefficient of c assumes the value 2 (Newton's value). The population P_j in Eq. (42.7) (and thus in Eq. (42.5)) may then be interpreted as mass m and the constant G as energy E . Also in this framework, the rank R_j may assume a cost interpretation. It should be noted that Isard (1971) also interpreted Einstein's law in spatial economic theory; however, in his view, the variable c represents flux or movement rather than a relative cost (or benefit) factor.

All in all, the compatibility between models (42.4) and (42.8) is fascinating. It seems that – by dealing with the complex network of masses (planets, cities, population, etc.) – the constancy of Newton's and Einstein's law persists, being captured by the SIM (42.4) and consequently by its general spatial formulation (42.1) or by its particular form, the rank-size rule (42.5).

From the theoretical viewpoint, as previously anticipated, the SIM is not only a “simple” model describing the spatial interaction between masses but is also the equilibrium solution of an entropy maximization approach. In addition, the SIM can be analytically linked to microeconomic theory, by means of the logit model, which emerged from random utility maximization (McFadden 1974). Compatibility between the SIM and the logit model has been demonstrated (Anas 1983; Sen and Smith 1995; Nijkamp and Reggiani 1992). The SIM can be interpreted in a behavioral context with an economic meaning, by considering the SIM to be an aggregate model of human behavior.

In summary, the above considerations highlight:

- The simplicity of the SIM
- The “constancy” of the SIM, with reference also to other disciplines and related laws, such as Newton's and Einstein's law
- The theoretical strength of the SIM, being connected to entropy maximization and to microeconomic theory

The SIM seems, therefore, to be the most simple and suitable model able to map out the static complexity of a network, from different angles (aggregate/disaggregate level) and from different spatial scales, by dealing also with a great number of origin/destination nodes. The issue of the SIM as an equilibrium “state” in a complex network evolution is examined next in Sect. 42.4.

42.4 Dynamic Complexity and Models

42.4.1 Simple Models vs. Dynamic Complexity

In this section we show that, by considering the dynamic setting of a spatial economic system, we find the same “constancy” of the spatial interaction form and that, in

particular, the SIM reflected in Eq. (42.1) – and hence the related models (a), (b), and (d) (in Sect. 42.3) – represents the steady state of network evolution.

By considering a *dynamic framework*, we have to keep in mind that the only mathematical instruments available which are able to model dynamic (un)stable and complicated patterns are the difference or differential equations. The most simple – and interesting – dynamic model is certainly the May model, usually called the “logistic” (or Pearl-Verhulst) map, namely, the nonlinear first-order difference Eq., which reads as follows (Gandolfo 1996):

$$x_{t+1} = ax_t(1 - x_t) \quad x \in [0, 1] \quad a \in [0, 4] \quad (42.9)$$

Equation (42.9) originally stems from biology: the time-dependent value x_t represents the observation of the variable x (biological population) at time t , and the parameter a represents the growth parameter that reflects the maximum per capita rate of x_t . May’s logistic Eq. (42.9) is a nonlinear model, which is very simple in its formulation, since it contains only one variable (x) and only one parameter (a). However, it can show chaotic and irregular movements, according to particular values of the parameter a and initial conditions. More specifically, cyclical behavior for the values $3 < a \leq 3.824\dots$ or unstable/chaotic movements for $3.824\dots < a \leq 4$ occur. At the bifurcation value $a = 3.824\dots$, a period of cycle 3 appears, giving rise – according to Li and Yorke’s theorem “Period Three Implies Chaos” – to the chaotic situation, where an uncountable number of aperiodic and periodic trajectories occur. It should be noted here that we define chaotic systems as the deterministic, nonlinear, dynamic systems which are able to produce *complex motions* of such a nature that sometimes seem completely random (Gandolfo 1996; Nijkamp and Reggiani 1992). Thus, the dynamic complexity previously defined, with its inherent impossibility to predict, is a clear feature of the chaotic systems.

Consequently, May’s logistic Eq. (42.9) turns out to be a fundamental example of “dynamic complexity,” emerging from a “noncomplicated” network. In other words, static complexity according to Casti’s definition is not satisfied here, showing that the various measures of static complexity are not necessary conditions for reaching dynamic complexity. Also disorganized complexity in the spirit of Weaver is not satisfied here, since Eq. (42.9) does not deal with a large number of variables (see Sect. 42.2). We can then conclude that May’s model is the first “simple” example of dynamic complexity.

In May’s formulation (42.9), x_t varies between 0 and 1, and thus, it may denote – in the spatial economics field – the dynamic probability of choosing a certain discrete alternative (transport mode, market product, etc.).

In contrast, the differential version of Eq. (42.9), that is, the following logistic equation in continuous time:

$$\dot{x} = bx(1 - x) \quad (42.10)$$

does not lead to any type of instability, independently of the values of the parameter b , as established by the Poincaré-Bendixson theorem. In fact, according to this

theorem, chaotic behavior can only arise in continuous dynamic systems with three or more dimensions (for a review, see Nijkamp and Reggiani 1992). Therefore, the “ability” of the analyst to correctly interpret and model the spatial system under investigation is a crucial issue. This is because the choice of the difference vs. the differential equation in one dimension leads to completely different dynamic trajectories in the presence of high values of the parameters a and b , that is, unstable vs. stable trajectories, respectively, with clear implications for prediction purposes.

A final reflection concerns the relevance – in practical terms – of the parameter a in Eq. (42.9), since the related high values, in particular values greater than three, induce cyclical/chaotic behavior. In other words, one might argue whether these high values of a are common in the dynamics of spatial economic systems. The answer is that high values of a – which are connected to the growth rates of the variable x at time t – might occur either in those systems which are characterized by fast dynamics, such as the financial and Internet networks, or in particular space-temporal windows of “less fast” systems, such as the traffic or technological networks. Consequently, the use of May’s logistic model (42.9) in detecting complex dynamic behavior seems unsuitable for the systems which display slow dynamics, such as the demography and physical infrastructure system, where the growth parameter a usually does not assume high values. However, this possibility might occur in some links or nodes of these slow systems in a certain time interval. In this case, an interesting issue is the dynamic relationship between the whole stable system and the corridor/center which is unstable: will the stable system be able to stabilize the unstable subsystem, or will it be destabilized by this unstable area? There are many examples in this respect, for example, a train crash which can destabilize the whole rail network and a terrorist attack on a central node. Analytically, it seems that the destabilization of the whole system might occur under particular interrelated conditions of the carrying capacities and parameter values (Nijkamp and Reggiani 1992, 1998).

The analysis of the relationship “stable vs. unstable system” implies an enlargement of our Eq. (42.9) to two or more dimensions. If x_t represents, for example, the dynamic production of a peripheral area, we need an additional variable y_t expressing the dynamic production of the metropolitan area, strictly linked to x_t , as well as additional terms expressing the dynamic interaction between x_t and y_t (and vice versa). Examples of dynamic Eqs in two and more dimensions will be provided in the subsequent Sect. 42.4.2.

Having said this, it is worth returning to the issue of the methodological “strength” of the SIM also in a dynamic setting. In this context, it can be shown that the logistic Eq. (42.9) is the dynamic version of a binary logit model and thus of a SIM (since a logit model is compatible with a SIM), under the condition that the utility function of x_t increases linearly with time through the fixed parameter a . In summary, a dynamic SIM might also exhibit complex behavior, since it is strictly connected to May’s equation of type (42.9) (Nijkamp and Reggiani 1992). In addition, the SIM appears to be the equilibrium solution of a dynamic entropy maximization approach, thus reinforcing the argument that a random complex network (like that expressed by a dynamic entropy) shows the SIM to be a simple model in its equilibrium.

It is then interesting to examine a multiple-choice situation, when the number of dynamic variables increases ([Sect. 42.4.2](#)).

42.4.2 Less Simple Models vs. Dynamic Complexity

In a multiple-choice dynamic situation, the logistic Eq. (42.9) shows the addition of the interacting terms, in the form of an ecologically based model, like the well-known prey–predator, competing, or symbiosis system. Let us consider, for example, a general competing system in two dimensions (in discrete time):

$$\begin{aligned}x_{t+1} &= x_t(H - hx_t - ey_t) \\y_{t+1} &= y_t(V - fx_t - vy_t)\end{aligned}\tag{42.11}$$

where x_t and y_t represent, respectively, the values of the variables x and y at time t ; H , V , h , and v are related to the endogenous variable dynamics of each corresponding variable; and the coefficients e and f reflect the interaction between the two dynamic variables x and y . System (42.11) is a general formulation which clearly interrelates two logistics of type Eq. (42.9); depending on the signs of the parameters, either the well-known prey–predator model developed by Lotka and Volterra, or the competition/symbiosis model emerging from Eq. (42.11) ([Gandolfo 1996](#); [Nijkamp and Reggiani 1992, 1998](#)). Since system (42.11) is expressed in discrete time, unstable and chaotic/unpredictable trajectories may emerge, depending on the values of the parameters and initial conditions, according to the Poincaré-Bendixson theorem, previously outlined.

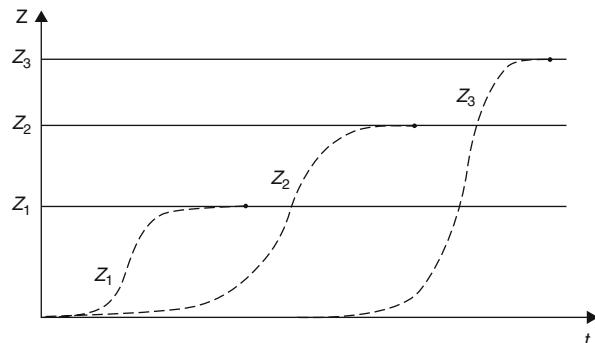
System (42.11) has frequently been utilized in spatial economic analysis as an “epidemic” model for describing technological innovation diffusion, urban growth, etc. (see [Batty 2005](#); [Fischer and Leung 2001](#); [Nijkamp and Reggiani 1998](#)). Interestingly, like the logistic Eq. (42.9), system (42.11) is very simple, although it can show complex and unpredictable patterns. In fact, by varying the parameters, simulation experiments concerning networks of form Eq. (42.11) exhibit a wide spectrum of ordered, irregular, and complex behavior.

From the empirical viewpoint, such results raise the problem of the prediction and control of complex systems and, hence, the necessity to collecting quasi-dynamic or panel data on spatial networks, in order to test the model coefficients and mostly their speed of change.

A generalization of Eq. (42.11) is the niche model, which expresses the phenomenon of interspecies competition and dynamic resource utilization. The niche model can also be interpreted in an economic framework, by considering the interaction between species as production functions ([Nijkamp and Reggiani 1998](#)). Formally, the logically defined niche system (in continuous terms) is:

$$\dot{z}_i = z_i(Z_i - \sum_{j=1}^N d_{ij}z_j)\tag{42.12}$$

Fig. 42.1 The visualization of Eq. (42.12) for three populations (z_1), (z_2), and (z_3). Here the logistic niche (z_1) is occupied successively by a new niche of increasing effectiveness (z_2); analogously, the logistic niche (z_2) is occupied successively by a new niche of increasing effectiveness (z_3)



where z_i is the population of species i (e.g., transport mode or market product $i = 1, 2, \dots, N$), \dot{z}_i is the rate of change of z over time, Z_i is the carrying capacity for species i , and the coefficients d_{ij} are the interaction/competition coefficients measuring the niche overlapping. The logistic niches can be visualized in Fig. 42.1 for three populations (z_1 , z_2 , and z_3). A positive network evolution occurs when a new species (z_2) replaces, in the short or long run, the old one (z_1), by exploiting new network capacities; analogously, the logistic niche (z_2) is occupied successively by a new niche of increasing effectiveness (z_3) (Nicolis and Prigogine 1977). The dynamic processes of the substitution/introduction of innovations in technology, transport, and market goods, as well as the development of new patterns of urban growth, can be modeled by means of the niche chain (Eq. (42.12)).

Interestingly, system (42.12), even though expressed in continuous time, may exhibit chaotic and complex behavior, for $i > 3$, that is, in the presence of three or more species of population, again according to the Poincaré-Bendixson theorem. It should be noted that the capacities Z_i and the coefficients d_{ij} may also embed dynamic functions, by modeling what are called “multilayer niche dynamics,” where the interaction exists not only among niche species but also among the niche capacities and/or the niche growth rates. However, by increasing the number of dynamic variables and parameters in model (42.12), very often the model fails in its analytical potential, by showing that a multilayer complicated model can hardly display complex behavior.

In other words, if we increase the analytical complexity of the network (by increasing, e.g., the number of variables, the multilayer/multilevel configurations) in order to replicate the real world, surprisingly the complex motions very often tend to explode or disappear: we reach the stage of “mathematical undecidability,” where no suitable information can be extracted (Reggiani 2004). In short, an inverse relationship between analytical complexity and dynamic complexity seems to emerge.

42.4.3 Concluding Remarks

In this section, we have pointed out that the SIM represents the steady state of network evolution, by maximizing a dynamic entropy. Moreover, the dynamic

version of the logit model (compatible with the SIM) leads, under particular assumptions, to the well-known (May/Pearl-Verhulst) logistic function, which can lead to unstable and unpredictable behavior of a chaotic type. In other words, the simple models (a)–(d), and in particular SIMs, appear to be the fundamental (conceptual and operational) instruments able to *decode* the complexity of the space-time phenomena concerned (Reggiani 2004; Reggiani and Nijkamp 2009).

From the methodological viewpoint, the lack of predictability of future events – for complicated systems – is still the main issue in the research concerning complexity in spatial economic analysis. A contribution in this respect might be provided by the exploration of network analysis. It is interesting to note the following: if the search for a “hidden” order/simplicity seems to have governed the scientific arena in spatial economics as an instrument able to “decode” and harness complexity, then network analysis, which aims to study complex network representations of physical, biological, and social phenomena, again reveals these “simplicity laws.” This issue will be discussed next in [Sect. 42.5](#).

42.5 Complexity and Network Analysis

42.5.1 Preface

In [Sect. 42.2](#), we indicated that complex systems evolve in different ways, depending on the type of interdependencies among the components. Thus, connectivity, that is, the ability to make and maintain a connection between two or more points in a spatial system, is one of the essential elements that characterizes complex networks. The connectivity issue has been strongly emphasized in recent years, especially in social network analysis, with consequent impacts and developments in other fields.

“A social network is a set of actors (individuals or social groups) and relationships of different kinds (friendship, kinship, status, sexual, business or political) among them” (Boccaletti et al. 2006, p. 251). In this chapter, we do not deal with social networks, since our attention is focused on complexity and spatial networks. However, it is useful to recall that some fundamental concepts and tools, which are now used in network analysis, such as connectivity, node centrality, and clustering index, have their origin in sociometry (Boccaletti et al. 2006). For a review on social network analysis, see, among others, Scott (2000). It is also interesting to recall here that the fast development of communication systems (Internet, cellular phones, etc.) has created new (virtual) forms of social contacts and cooperation, which can be modeled by means of network analysis (see, e.g., the analysis of community structures in the context of R&D Cooperation in Europe, by Barber et al. 2011).

In spatial economics, the connectivity concept has hardly been formalized, since it has been encapsulated in the strength of the network interaction and thus embedded in the values of the variables concerning the models previously mentioned, essentially in the cost matrix c_{ij} . In this context, connectivity has been

strictly linked to the concept of accessibility, since accessibility weights the network connectivity structure – embedded in the cost matrix c_{ij} – by means of the socioeconomic activities in j (for a review on the link between accessibility and connectivity, see Reggiani 2012). Connectivity is now receiving more attention, thanks to the popularity of social and network analysis. An interesting related issue is the relevance of the topological network structure, particularly considering the often conflicting relationship between distance/cost and topology. For example, two spatially close neighborhoods may not display any significant interaction if they are separated by a strong barrier (e.g., a highway).

In addition, networks often show common behavior, based on their topological characteristics; consequently, the identification of the network architecture/topology cannot be ignored without missing a crucial ingredient of the complex phenomena concerned (Vega-Redondo 2007). The topology issue implies a focus on the network configuration and its properties, in order to analyze the related impact on the behavioral dynamics of the network itself.

Starting from this issue, that is, the relevance of connectivity and topology in complex networks, in this section we show how network analysis deals with it, in particular by highlighting two focal network models (random and scale-free networks) which are strictly linked to the aforementioned models (see Sect. 42.3) conceived and applied in spatial economic analysis.

42.5.2 Simple Network Models: Random and Scale-Free Networks

In network science, a rigorous framework for the description and analysis of networks is found in graph theory. We can refer, first, to 1736, which marked the birth of this discipline, when Leonhard Euler published the solution to the Königsberg problem, and then to the 1920s, which witnessed the early beginnings of social network analysis that focuses on the complex relationships between social entities. In the last few decades, there has been renewed interest in the study of complex networks by means of graph theory, thanks essentially to the works by Watts and Strogatz on small-world networks and by Barabási and his group on scale-free networks (for a comprehensive review, see Boccaletti et al. 2006).

It is interesting to observe how complexity is defined in network science. For example, Caldarelli and Vespignani (2007), p.15 argue that “a definition of complexity may involve two main features: (i) the system exhibits complications and heterogeneity that extend virtually on all scales allowed by the physical size of the system; (ii) these features are the spontaneous outcome of the interactions among the many constituent units of the system, i.e. we are in the presence of an emergent phenomenon.” These authors add that examples of this are the WWW, the Internet, the airline airport networks, and all the social and biological networks which grow in time by following complicated dynamic rules and without global supervision. Moreover, “All these networks are self-organizing systems, which at the end of the evolution show an emergent architecture with unexpected properties and regularities” (Caldarelli and Vespignani 2007, p. 15).

Two new interesting characteristics of complex networks were introduced by Caldarelli and Vespignani: (a) emergence and (b) self-organization. Complexity is “decoded” here by means of the “emergence” level, which results from the self-organized, spontaneous, process coming out from the complicated or complex interaction of the units at the lower levels. Caldarelli and Vespignani continue their discussion on complexity by pointing out that heavy tails and heterogeneity appear to be the common features of a large number of these complex networks. In other words, the emergence concept highlights a final state of the complex system which can be identified and mapped out, thanks to its “regular” properties.

In particular, the network topology seems to be crucial in determining the emergence of this “collective” dynamic behavior (such as synchronization of activities, habits, fashions, ideas) or in governing the main features of relevant processes (such as the spreading of information, epidemics, rumors, new ideas) (Boccaletti et al. 2006). In this context, one of the most surprising findings in social network analysis is that real networks behave very differently from conventional hypotheses about them. Traditionally, real networks were conceived to have a majority of nodes with about the same number of connections around an average. These are called “random networks” (Erdős and Rènyi 1959), which display homogeneous, diffuse patterns, without cluster characteristics. In a random network with n ($n = 1, \dots, N$) nodes and k links, the degree (number of links k per node n) distribution $P(k)$ is well approximated by a Poisson distribution, as follows:

$$P(k) = e^{-\langle k \rangle} \frac{\langle k \rangle^k}{k!} \quad (42.13)$$

where $P(k)$ is the probability that a node n chosen uniformly at random has degree k and $\langle k \rangle$ is the average degree. Eq. (42.13) identifies the random network as a homogeneous network. Homogeneity, in the interaction structure, means that all the nodes are topologically equivalent, and thus, each link is present with equal probability.

However, a variety of different social, natural, economic, and technological networks significantly deviate from the Poisson distribution (42.13), since they show high clustering characteristics, degree correlations, and the presence of motifs (patterns of interconnections) and community structures (subgroups or clusters tightly connected). All these common features make these real networks completely different from random graphs: they display fat-tailed shapes in their degree distributions (Boccaletti et al. 2006). In summary: in all these situations, the associated graph presents a universal feature of some elements with many connections (hubs) and many with only a few. This is expressed by a power-form-shaped degree distribution, as follows:

$$P(k) \sim k^{-\gamma} \quad (42.14)$$

where γ is the degree exponent and \sim indicates “proportional to.”

The networks which display a power form of type (Eq. (42.14)) have been called “scale-free networks” by Barabási and his group, because power distributions display the same functional characteristics at all scales. The value of the exponent γ depends on the attributes of the single systems and is crucial to detect the exact network topology. As Barabási and Oltvai (2004) point out, the value of γ determines many properties of the system. The smaller the value of γ , the more important the role of the hubs is in the network. In particular, for $2 < \gamma < 3$, there is a hierarchy of hubs, with the most connected hub being in contact with a small fraction of all nodes, and for $\gamma = 2$, a hub-and-spoke network emerges, with the largest hub being in contact with a large fraction of all nodes. In general, the properties of scale-free networks are valid only for $\gamma < 3$. For $\gamma \geq 3$, the hubs are not relevant, and in many respects, the scale-free network behaves like a random one.

Clear empirical examples concerning random networks and scale-free networks can be found in spatial economic systems. For example, in air transport, random networks are useful to map point-to-point connections, as is the case for low-cost airlines. In the ideal point-to-point network, all airports are connected to each other so that passengers can fly from one airport to any other directly without stopping in any hub to change flights. The same applies to any other type of network which can be seen as a homogeneous system which gives accessibility to the majority of the nodes in the same way.

In contrast, the Internet, the WWW, the high-speed train, the air transport system with full-service carriers, scientific coauthors, company directors, and any other socioeconomic network typified by what is called a hub-and-spoke structure (where central nodes (hubs) have a high number of links (spokes) to the majority of the other nodes), all exhibit this clustering nature, which is also reflected in the associated socioeconomic activities. It should also be noted that the scale-free network was introduced by Barabási in order to incorporate two mechanisms upon which many real networks have proved to be based: *growth* and *preferential attachment*. The former indicates the dynamic character of networks, which grow by the addition of new nodes and new vertices; the latter explains how new nodes enter the network, namely, by connecting themselves to the nodes which have the highest number of links.

42.5.3 Concluding Remarks

In the previous section, we briefly delineated two main types of complex networks, namely, random and scale-free networks, both of which have recently received a great deal of attention in the literature on complex networks, especially in social network analysis.

In this framework, the value of the coefficient γ in the connectivity degree distribution (42.14) appears to be crucial in the identification of the random/scale-free network and thus of related homogeneity/heterogeneity characteristics of the associated network. Other network topology indicators, like centrality, closeness, betweenness, and cluster coefficients, can also help in the network pattern recognition

(for a definition of these indicators, see, among others, Barthélémy 2010; Boccaletti et al. 2006). In addition, if we want to detect the network configuration (random vs. scale-free), we need to understand to what extent these networks are concentrated, because the existence of hubs implies a high degree of concentration. Thus, the network concentration indices, such as the Gini concentration index, the Freeman centrality index, and the entropy indicator, can be useful in this respect.

In summary, these interesting studies on random vs. scale-free networks have revealed, on the one hand, the importance of the topological characteristics in a complex network for the detection of homogeneity/heterogeneity features and, on the other hand, that these topological characteristics are captured by very simple indicators, as we can see by examining Eq. (42.13), identifying random networks, and Eq. (42.14), identifying scale-free networks. Basically, the essential element which leads to this topological diversification (random vs. scale-free network) is the type of connectivity. This issue will be examined in the next section, also with reference to spatial economic analysis.

42.6 Spatial Economics and Network Analysis: Connectivity, Emergence, and Resilience

Our debate on complexity in Sect. 42.3 mainly focused on dynamic complexity. In particular, the complex network's characteristic of high levels of interdependence through nonlinearities drew attention to the fundamental feature that the outcome is not obvious from the simple building blocks. In Sect. 42.4 we saw how complexity can be conceived as the result of a complex (and possibly evolving) network of connections among the different units involved. Hence, connectivity assumes a fundamental role in detecting complexity.

Interestingly, if we carry out a dual analysis, that is, spatial economic analysis vs. network analysis, we find that the homogeneity and heterogeneity of the economic centers in spatial economic analysis fit the homogeneity and heterogeneity of the topological structures (random vs. scale-free networks) in network analysis. For example, the rank-size rule and Zipf's law, which model the urban hierarchy (see Sect. 42.3), were conceived in a historical period (the 1950s) when the physical and virtual connectivity between cities was not so evident and strong as it is today. However, the cities are connected, and in fact, it has been recently demonstrated that the value of the coefficient q , which can be interpreted as the elasticity parameter in the rank-size rule (42.5), thus identifying the type of hierarchical spatial structures in spatial economics, is strictly related to the degree exponent γ , which emerges from the connectivity distribution (42.14) in the associated graph. In other words, the coefficients q and γ appear to be two sides of the same coin; more precisely, the q -coefficient identifies the population (in)equalities from the viewpoint of spatial economics, and at the same time, it is linked to γ by this simple analytical transformation (Adamic 2000):

$$\gamma = 1 + (1/q) \quad (42.15)$$

which can be crucial in the identification of the associated random/scale-free network and thus of the homogeneity/heterogeneity network connectivity characteristics. In particular, by considering Eq. (42.15), we can find the following correspondence:

- For $q \leq 0.5$ (urban homogeneity), $\gamma \geq 3$ (random network) emerges.
- For $q > 0.5$ (urban heterogeneity), $\gamma < 3$ (scale-free network) emerges.

This connectivity interpretation of the rank-size coefficient q reinforces the argument on the relevance of the connectivity element in the spatial economic networks, which was somehow neglected in the models based on the SIM and also mentioned earlier in Sect. 42.3. As a consequence, the emergence concept assumes a new “network” meaning. As previously anticipated in Sect. 42.4.2, the emergent phenomenon is the “state” of the network expressing collective behavior, that is, a self-organized structure which is the result of the continuous dynamic interplay between the macro- and the microelements of a network. Emergence tells us that an economic system of interacting agents (like traffic commuters or traders in a financial market) can spontaneously develop collective properties that are not at all obvious from our knowledge of each of the agents individually. Emergence signifies order despite change (Reggiani 2004). However, this interplay between the dynamic behavior of the agents at the microlevel is only possible by means of connectivity. The emergent mesostructures can, therefore, also be identified as random, scale-free, or intermediate networks. For example, if we consider the logistic niche, as in Eq. (42.12) and Fig. 42.1, the envelopes z_1 , z_2 , and z_3 , representing the emerging network structures resulting from the interaction between the dynamic behavior of SIM structures at the microlevel, can also be classified as random/scale-free or intermediate networks, according to their connectivity structures.

Consequently, the connectivity and emergence concepts emphasize the evolutionary aspect of organized complexity, in contrast to disorganized complexity (Sect. 42.2). These two ingredients reinforce the perspective of order in complexity.

A final issue, related to the previous ones, which is worth mentioning, is the fragility/resilience aspect of a complex network. An important feature of the scale-free network highlighted by Barabási and his coauthors is a high degree of robustness in the face of accidental node failures. In other words, in the case of a random attack on nodes, the scale-free network will show high resistance, because a random attack will probably damage nodes that have only a few connections (which are the majority). In contrast, random networks are weak against a random attack which will cause a rupture of the network.

However, in the case of an oriented attack against the hubs, the network will easily be fragmented, because of the high connectivity of the hubs with the majority of the nodes. Consequently, we might also talk of the “vulnerability/permeability” of the scale-free network in its hubs: if certain information or a virus is dispersed in the hubs, it is diffused all over the network, if the connective configuration is completely accessible. Accessibility then turns out to be a driving force for the formation of the scale-free network and the related dynamic functionality. Therefore, the identification of the random/scale-free characteristics, together with the

associated accessibility patterns, appears to be essential for understanding the dynamics of network function and behavior in the light of policy/planning interventions (Reggiani 2012).

The issue of the relationship between network stability and complexity thus opens new perspectives, essentially based on the concepts of connectivity, emergence, and fragility/resilience. In general, a dynamic fragile system can be defined as a system that will tend to collapse under perturbations to its parameters or population values (as in the case of the scale-free networks attacked in its hubs). In this framework, the concept of resilience appears to offer interesting ground for investigating the stability structure of a complex network. Resilience refers to the capacity of a system to retain its organizational structure following the perturbation of some state variable from a given value, but not only, since resilience also reflects the capacity of the network to adapt itself to new states; thus, evolution is formed by the switch of these resilient networks from one equilibrium state to another (for a review on resilience applied to spatial systems, see, among others, Rose 2009).

Resilience reveals a framework that goes beyond the usual stability concept, since, in principle, a complex system can be unstable but resilient (Reggiani 2004). In other words, resilience can overcome the conventional debate “unstable/stable node vs. stable/unstable network” – outlined in Sect. 42.2 – by allowing unstable paths toward different equilibrium states in the complex network (as in the case of scale-free networks, if attacked in the majority of the nodes which are not the hubs).

42.7 Conclusions

This chapter has aimed to review briefly the complex relationship “complexity and spatial networks.” The argument is so vast that it is impossible to tackle this issue from all the perspectives of analysis. We focused, therefore, only on the main models used in the regional/spatial economic literature as well as in network analysis, in order to compare and investigate similarities and differences.

The following main conclusions can be drawn: a formal correspondence between the rank-size rule (42.5) and network connectivity analysis expressed by Eq. (42.14) does exist, thanks to the behavioral interpretation of the q -coefficient and to its related connectivity γ -coefficient. Thus, the rank-size model (42.5), and hence the SIM (42.1), is able to (i) grasp the homogeneity/heterogeneity of the network concerned (at an aggregate level, by means of its q -coefficient, expressed in Eq. (42.5)) and (ii) represent the associated connectivity infrastructure or socio-economic network/constellation of complex spaces, by means of its associated γ -coefficient, as in Eqs. (42.14) and (42.15).

From the spatial economic viewpoint, Eq. (42.5) is conventionally formulated according to a power form. However, it has been demonstrated that Eq. (42.5) can theoretically embed different functional forms, such as exponential and lognormal, all of them capable of capturing the socioeconomic spatial characteristics of the network under analysis. In summary, the power form in Eq. (42.5) with a q -coefficient > 0.5 is suitable to detect the presence of agglomeration economies,

that is, hierarchies in the spatial economic structure of the variable concerned. As is well known, Eq. (42.5) could also map out different variables other than population, such as GDP and inflows. Thus, both the simple rank-size model (42.5) and the SIM (42.1) certainly represent a useful instrument in grasping the “emergent” features of the spatial economic networks.

From the network viewpoint, formulation (42.14) has reinforced the argument that complex phenomena can exhibit unexpected similarities, as well as increasing the interest in researching what is called the “heavy tail” in the probability distribution of a certain quantity, and thus heterogeneity in the number of connections per node.

Thus, analytically, the power form seems to be “ubiquitous” from different perspectives, expressing the aforementioned inequality characteristics. However, some caution is necessary here, with reference to (a) the estimation analysis of the coefficients which should be statistically correct and (b) the theory behind formulation (42.14), which so far seems to be derived from empirical experiments.

This latter issue necessitates some reflections on data. An alternative way of detecting complex behavior is the use of techniques able to extrapolate, from data, nonlinear network interactions. It should be noted that in chaos and complexity analysis (Sect. 42.4), many applications lack empirical content. A solution could be the adoption of techniques generally used for proving the existence of chaotic behavior – and thus the inherent dynamic complexity – in panel data. Detecting complexity from data requires the use of conventional established techniques, such as the Brock-Dechert-Scheinkman statistic, the method of the largest Lyapunov exponent, or the artificial neural network tool, belonging to the biocomputing models. In this context, it is interesting to mention works that address the compatibility between artificial neural networks and SIMs (e.g., Reggiani 2004).

Finally, some simulation tools, for example, cellular automata (based on a fixed spatial framework) and agent-based modeling (where agents can be mobile with respect to space), should be mentioned here, on account of their potential in detecting emerging patterns. Cellular automata and agent-based modeling are complementary modeling strategies. They can be integrated into a common geographic automata system where some agents are fixed, while others are mobile (Batty 2005).

New research paths should then consider, in a multidisciplinary way, a very rich agenda, which mostly tries to join these two disciplines, spatial economics and network science, from all the perspectives: theory, methodology, empirics, and policy analysis.

Currently, network analysis appears to be extremely full of new contributions which aim to deepen the first findings of Barabási and his coauthors: the elaboration of new network metrics, as well as of new dimensions of shock propagation (e.g., depth, width, strength), has recently come to the fore. All this shows the efforts that have been made in approaching spatial science and its modeling. On the other hand, an increasing number of studies in regional science are now adopting the models of network analysis in order to analyze the space-time dynamics of economic phenomena.

This endeavor involves a synthesis of knowledge from the different scientific traditions, where the traditional concept of prediction needs to be revisited, in the presence of a complex network sensitive to initial conditions and perturbations. Mostly, new developments in theory are necessary, for example, new theories of stochastic dynamics or path-dependent dynamics, in the presence of large networks, given the current rich amount of data available in telecommunications and the powerful computation tools.

Finally, a further effort, in the form of a blend of advanced theories and approaches belonging to regional economics and complex network theory, is required to bridge the gap between science and policy, in order to provide an integrated framework able to manage the multilayer-multilevel complex spatial networks also from an operational viewpoint.

Acknowledgment The author wishes to thank two referees for their valuable comments.

References

- Adamic LA (2000) Zipf, power-laws, and pareto – a ranking tutorial. Retrieved 18 May 2012 from: <http://www.hpl.hp.com/research/idl/papers/ranking/ranking.html>
- Anas A (1983) Discrete choice theory, information theory and the multinomial logit and gravity models. *Transport Res B* 17(1):13–23
- Axelrod A, Cohen MD (2000) Harnessing complexity. Basic Books, New York
- Barabási AL, Oltvai ZN (2004) Networks biology: understanding the cell's functional organization. *Nat Rev Genet* 5(2):101–113
- Barber MJ, Fischer MM, Scherngell T (2011) The community structure of research and development cooperation in Europe: evidence from a social network perspective. *Geogr Anal* 43(4):415–432
- Barthélemy M (2010) Spatial Networks. Retrieved 13 January 2013 from: <http://arxiv.org/pdf/1010.0302.pdf> (Published in 2011. *Phys Rep* 499:1–101)
- Batty M (2005) Cities and complexity: understanding cities with cellular automata, agent-based models, and fractals. MIT Press, Cambridge
- Batty M (2010) Space, scale, and scaling in entropy-maximising. *Geogr Anal* 4(1):395–421
- Boccaletti S, Latora V, Moreno Y, Chavez M, Hwang DU (2006) Complex networks: structure and dynamics. *Phys Rep* 424:175–308
- Caldarelli G, Vespignani A (2007) Large scale structure and dynamics of complex network. World Scientific Publishing, Singapore
- Casti J (1979) Connectivity, complexity and catastrophe in large scale systems. Wiley, Chichester
- Einstein A (1905) Ist die Trägheit eines Körpers von seinem Energieinhalt abhängig? (Does the inertia of a body depend upon its energy-content? *Annalen der Physik* 323 (13): 639–643. Retrieved 24 May 2012 from: <http://onlinelibrary.wiley.com/doi/10.1002/andp.19053231314/pdf>
- Erdős P, Rényi A (1959) On random graphs. I. *Publ. Math. Debrecen* 6: 290–297. Retrieved 24 May 2012 from: http://www.renyi.hu/~p_erdos/1959-11.pdf
- Fischer MM, Leung Y (2001) Geocomputational modelling. Springer, Berlin/Heidelberg/New York
- Fotheringham AS, O'Kelly ME (1989) Spatial interaction models. Formulations and applications. Kluwer, Dordrecht
- Gandolfo G (1996) Economic dynamics. Springer, Berlin/Heidelberg/New York
- Heylighen F (1996) What is complexity? Retrieved 11 March 2012 from: <http://pespmc1.vub.ac.be/COMPLEXI.html>.

- Isard W (1956) Location and space-economy. MIT Press, Cambridge
- Isard W (1971) On relativity theory and time-space models. Pap Reg Sci Assoc 26:7–24
- Krugman P (1994) Complex landscapes in economic geography. In: Reggiani A, Button K, Nijkamp P (eds) Planning models. Classics in planning. Edward Elgar, Cheltenham, pp 401–405
- McFadden D (1974) Conditional logit analysis of qualitative choice behaviour. In: Zarembka P (ed) Frontiers in econometrics. Academic, New York, pp 105–142
- Nicolis G, Prigogine I (1977) Self-organisation in non equilibrium systems. Wiley, New York
- Nijkamp P, Reggiani A (1992) Interaction, evolution and chaos in space. Springer, Berlin/Heidelberg/New York
- Nijkamp P, Reggiani A (1998) The economics of complex spatial systems. Elsevier, Amsterdam
- Reggiani A (2004) Evolutionary approaches to transport and spatial systems. In: Hensher DA, Button KJ, Haynes KE, Stopher PR (eds) Handbook of transport geography and spatial systems. Elsevier, Amsterdam, pp 237–252
- Reggiani A (2012) Accessibility, connectivity and resilience in complex networks. In: Geurst KT, Krizek KJ, Reggiani A (eds) Accessibility and transport planning. Edward Elgar, Cheltenham, pp 15–36
- Reggiani A, Nijkamp P (2009) Complexity and spatial networks. Springer, Berlin/Heidelberg/New York
- Rose A (2009) Economic resilience to disasters, CARRI Report No. 8, Community and Resilience Institute. Retrieved 11 March 2012 from: http://www.resilientus.org/library/Research_Report_8_Rose_1258138606.pdf
- Scott J (2000) Social network analysis. Sage, Newbury Park
- Sen A, Smith TE (1995) Gravity models of spatial interaction behavior. Springer, Berlin/Heidelberg/New York
- Simon H (1962) The architecture of complexity. Proc Am Philos Soc 106(6):467–482
- Vega-Redondo F (2007) Complex social networks. Cambridge University Press, Cambridge
- Weaver W (1948) Science and complexity. Am Sci 36:536–544
- Wilson A (1970) Entropy in urban and regional modelling. Pion, London
- Zipf GK (1949) Human behaviour and the principle of least effort. Addison-Wesley Press, Cambridge

Market Areas and Competing Firms: History in Perspective

43

Folke Snickars

Contents

43.1	Introduction	834
43.2	Theoretical Background	836
43.3	Theoretical Modeling of Location Choices	839
43.4	Basic Modeling Principles and Assumptions	841
43.5	Experimenting with the Hotelling Model	843
43.6	Analysis of the Simulation Results	847
43.7	Conclusions	848
	References	849

Abstract

Location theory has traditionally been based on equilibrium concepts. Dynamics have been introduced mainly to ascertain whether there are paths leading to the equilibrium states. The modeling of dynamics has been very simple yet involving both locational changes and price changes. Notions of market areas and competition between firms have been at the core of location analysis. Although the classical location theory was developed in a regional context, the models have found a number of recent applications in urban analysis where interdependencies and dynamics are central elements. The theoretical contributions of Hotelling, Hoover, and Palander form cornerstones for the discussion in the current chapter. In this chapter, we will mainly dwell in the Hotelling tradition and use the theories of Hoover and Palander as introductory and complementary inputs. The chapter presents a series of behavioral models in the spirit of the classical Hotelling location game involving the spatial location

F. Snickars

Department of Urban Planning and the Environment, KTH Royal Institute of Technology,
Stockholm, Sweden
e-mail: folke.snickars@abe.kth.se

of suppliers (sellers) and consumers (customers) in an urban context. The models have been established within a cellular automata framework. The location models studied assume fixed prices. The location of sellers is determined by the relative accessibility to customers and the competition between sellers for customers. Using the techniques of cellular automata, a set of simulations will be performed to discuss equilibrium states of customer-seller systems. The discussion will serve to illustrate some elements of location theory under different levels of complexity.

43.1 Introduction

Some scientific articles have become classics in their field. This is the case with the theoretical contributions of Hotelling, Hoover, and Palander. Their works form one of the cornerstones for the discussion in the current chapter. We will focus mainly on the Hotelling tradition and use the theories of Hoover and Palander as introductory and complementary inputs. Techniques of cellular automata are used to investigate how fundamental principles of dynamics and evolution which replicate the theoretical results of locational analysis can be applied also in more complex spatial arrangements. This type of analysis has been demonstrated useful in studying a wide range of dynamic systems, including spatial urban systems (White and Engelen 1997; Semboloni 2000), political systems (Downs 1957), and innovation systems (see, e.g., Rasmussen 1989; Leydersdorff 2002).

It is common to consider industrial location within conventional general equilibrium theory, in which everything is assumed to happen at one point in space. Two fundamental questions have to be distinguished: Where will production take place? And given the place of production, the competitive conditions, factory costs, and transportation rates, how does price affect the extent of the area in which a certain producer can sell his goods?

One of the fundamental research issues in location theory is the boundary of the market areas of spatially located firms. The simple case of two firms making the same product for linear markets, where consumers are uniformly distributed on a line or along a street, was developed among others by Palander (1935).

Palander argues that the price charged at a certain location (the delivered price) is the plant cost measured as the price charged for the product at source plus the necessary cost of transportation to the fixed location from the plant. This is illustrated in Fig. 43.1. The boundary of the two markets will be the point where the delivered price from both producers is equal. This is the point where customers will be indifferent as to which firm they buy from. The size of the market area influences the profit. With the cost of production and profit per unit of output given, total profits become a function of the distance from the plant that a firm can extend its market.

The assumption of perfect competition was also used by Hoover (1938, 1948). Using the same setting as developed by Palander and introducing the condition of diminishing returns to scale in the production function of the firm, Hoover arrives at

Fig. 43.1 The classical result of Palander's market area theory

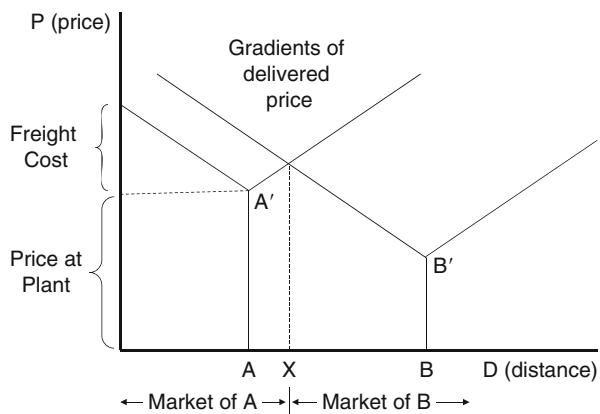
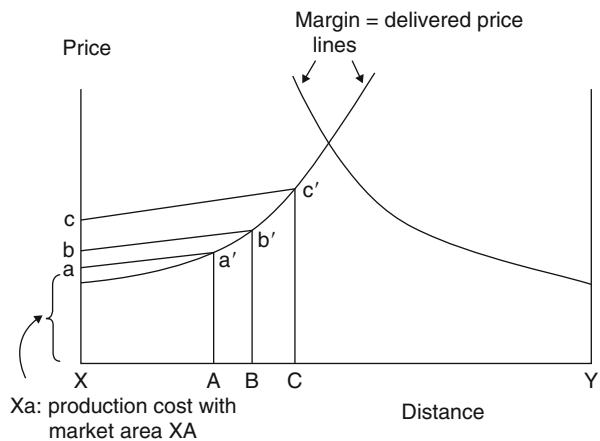


Fig. 43.2 The classical result of the Hoover location theory



the conclusion that in the absence of production cost differences, the best location will be at the point of minimum transport costs (see Fig. 43.2). Market area boundaries between different producers arise from areal variations in production costs and delivered prices.

In the early literature, market areas were separated from each other for firms having a fixed set of locations. The early location theorists found the problem of optimum location for the individual firm intractable. As soon as the interdependence of firms is accepted with the possibility that the action of one firm in locating itself can require the relocation of existing firms, the problem became too complex for mathematical formulations.

The third tradition from the classical location theory was the spatial competition analysis put forward by Hotelling (1929). That analysis differed in at least two respects from the other theorists. The first was that location was not given but a matter of competition. The other was that the analysis was dynamic or could at least be interpreted in a dynamic way. The analysis was aimed at providing a theory

which would apply to industries which would choose their location to compete for the best position.

The problem of location of interdependent firms in space was formulated and solved in strict theoretical terms by Koopmans and Beckmann (1957). They also showed that there was no price system which would sustain the equilibrium location in their quadratic assignment problem. It was shown by Heffley (1972) and Snickars (1978) that the reason for this complexity was the non-convexity of the quadratic assignment matrix. The nature of transportation costs implied that the matrices representing the assignment parameters, in essence, would never be diagonal dominant.

The following discussion will build on the tradition of Hotelling (1929) and combine it with game theory and the theory of cellular automata to address the question of spatial competition involving many competitors and several different demand schemes and competitive conditions. One reason for this choice of approach is that it is important to follow the dynamics of the system and thus to frame the classical theories in a modern theoretical and computational context.

43.2 Theoretical Background

Location theory has traditionally been based on equilibrium concepts. Dynamics have been introduced mainly to ascertain whether there are paths leading to the equilibrium states. The modeling of dynamics has been simple yet involving both locational changes and price changes. The current cellular automata framework is based on the assumption that complex spatially defined phenomena can be modeled by treating dynamics explicitly. Complexity will arise from the interaction among actors rather than from the behavioral assumptions for each actor. Cellular automata use a grid of cells to describe the spatial dimensions of the system and an incremental stepwise analysis of all cells to approximate the temporal dimension. This type of analysis has been demonstrated to be useful in a wide range of dynamic systems, but only recently the concepts are being applied to urban systems.

This contribution presents a series of behavioral models in the spirit of the classical Hotelling location game involving the spatial location of suppliers (sellers) and consumers (customers) in an urban systems context.

The location models in this chapter assume fixed prices at the factory gate. This does not preclude the possibility that these prices will vary among plants. Obviously, it will be beneficial for a firm to have lower production costs which will be reflected in the prices at the factory gate. The location of sellers is determined by the relative accessibility to customers and the competition between sellers for customers. The cellular automata approach used to investigate how fundamental principles of dynamics and evolution which replicate the theoretical results of locational analysis can be applied also in more complex spatial arrangements; see Fig. 43.3 for an example of two-dimensional location-theoretical results.

The dynamic simulations may also be used for predictive purposes to determine the equilibrium states of a customer-seller system at some future point in time. They

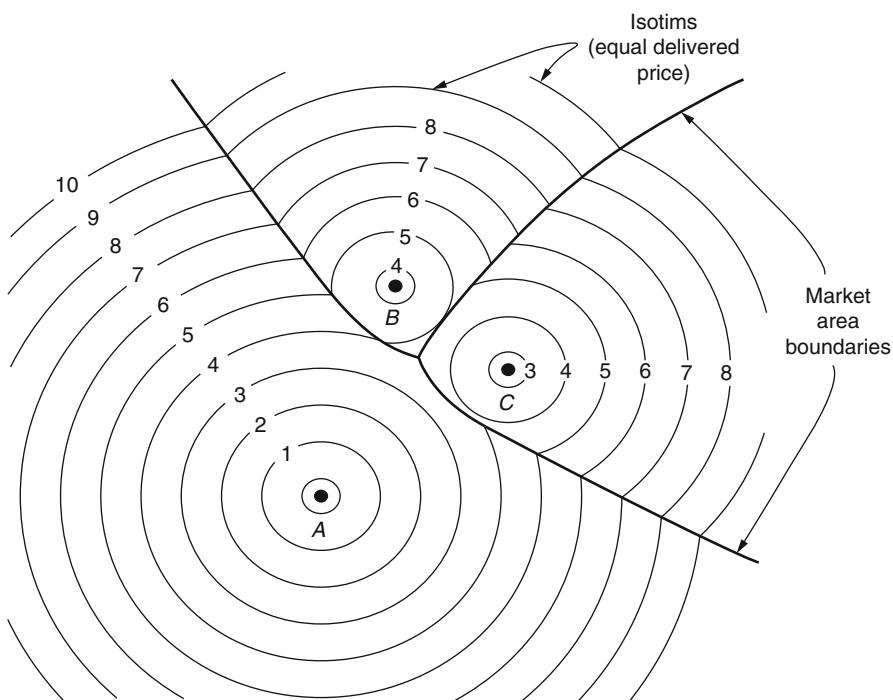


Fig. 43.3 Classical picture of the development of market areas in two dimensions

may also provide an insight into the dynamic behavior of customer-seller systems approaching the spatial complexity of real urban areas. Usually, these processes proceed concurrently and may, or may not, be directly related to one another. The speed of the processes will differ as will the spatial extension of impacts among agents involved in them.

It is difficult to provide a single approach for exploring the fundamental issues in the dynamics and spatial evolution of the urban system. A wide range of different approaches is emerging derived from the use of artificial intelligence, multi-agent-based models, cellular automata, network analysis, dynamic programming, queuing theory, game theory, stochastic simulation, and several other mathematical modeling techniques useful for urban analysis (see also Wegener 2004 and Batty 2008).

In the model treated, the simulation process operates on urban activities using sets of rules for spatial interaction among these activities, including environmental and other constraints. The set of urban activities (workplaces, residential districts, green areas, water surfaces) will vary in different investigations, depending upon the research goals, the level of aggregation required, and what interacting mechanisms of change among agents to be considered. The set of rules and constraints for spatial interaction may be more or less well defined and will be modeled using nonlinear system methods, differential equations, or procedural knowledge.

The processes will need to be spatially and temporally defined to model the inherent dynamics of multi-agent urban systems. In this chapter, these methods are used to develop a class of dynamic land-use models (see also Anas 1987; White and Engelen 1993; Roy and Snickars 1996, 1998; Wegener 2004; Batty 2008). Specifically, we will be examining a customer-seller problem in the spirit of the classical Hotelling location game to demonstrate these principles.

In the cellular automata approach adopted in the chapter, the spatial dimension is represented by a set of cells covering the region to be studied. These cells are often based on a regular grid, and while this is not a fundamental requirement, it does simplify computations involving distance and adjacency. The dynamic state of each cell is determined by its initial state and the dynamics of the states in neighborhood cells. Many of the properties of the cellular automata will be determined by the scope of the neighborhood and the rules of interaction among neighboring cells.

In addition to the rules of interaction among neighboring cells, there will be system-wide constraints and changes imposed by externalities outside the scope of the urban system itself. The constraints will often be employed to introduce different types of public policies to address the externalities caused by the interaction among urban agents. It is also possible to introduce agglomeration factors in the framework by increasing the attractiveness of clustering spatial location elements around already existing clusters.

The resulting modeling framework is rather simple and can be readily modified to add new interaction mechanisms while maintaining the same model structure. We still need to be concerned about the integrity of the modeling process to ensure that it adequately represents the interaction mechanisms we wish to include. The chapter is concerned with attempting to show, for the case of simple models, that we can obtain results consistent with classical theories. We will then extend the analysis with the help of our model to more complex situations where theoretical analysis will not be able to give closed-form results. A further ambition is to compare the equilibrium outcomes in a cellular automata framework to the ones derived from static equilibrium theory.

Similar land-use models based on cellular automata have been proposed by, for instance, Roy and Snickars (1996), White and Engelen (1993, 1997), Semboloni (2000), and Liu (2009). Most of them are intended for investigation of basic questions of emergent urban form rather than to provide simulations of the spatial development of particular cities. The contributions suggest different behavioral processes for the representation of urban activities. The common principles for all models are that the spatial structure of urban land-use is approximated by a regular grid of cells, each cell representing a single type of land use.

According to specified sets of transformation rules, the models convert cells from one state to another and so produce fractal or bi-fractal land-use structures for the urbanized area and for each land-use type. Transformation rules are generally simple and yet can produce highly organized, complex, evolving structures. The set of rules may be divided into those that permit an unused cell to be set to a certain state and those permitting cells to become locked in a particular state or to allow displacement of one land use by another. A set of weighting parameters is used to

represent the relative competitiveness of various land-use activities. These weighting parameters control the spatial and sectoral patterns of interaction. The choice of weighting factors is thus a part of the research investigation: to discover where and when particular values have significant influences on system behavior.

The simulation process begins from a predefined initial state for the land-use pattern and through a series of transformations that evolve into new states. The transformation rules for cells can be written quite generally, but they typically rely on the present state and the activities in neighboring cells to determine the subsequent state of the cell. The neighborhood concept may not necessarily be based on geographical proximity. It can also relate to linkages in relation to economic clusters of sectors of the economy or clusters in the sense of industrial districts. The model suggested by Roy and Snickars (1996) is more general in its approach to the simulation of urban system dynamics because it does not impose restrictions on the character of neighborhood bounds. Furthermore, it does not impose a predetermined development cycle where one land-use type has externally determined predominance to another.

43.3 Theoretical Modeling of Location Choices

As mentioned earlier, the model of location choice of sellers in an urban system is inspired by the classical article by Hotelling (1929). The topic has been treated extensively in later location theory work, see e.g., Gabszewich et al. (1986), Thisse et al. (1996), Fujita et al. (1999) and Nickel and Puerto (2005). A seminal contribution was made by d'Aspremont et al. (1979) where results were proved for general customer-seller configurations using methods from game theory. The Hotelling location analysis addressed the question of competition in space by letting sellers compete for customers both with their choice of location and with their product price. The question was what would be the equilibrium location and price configuration for different assumptions about customer behavior and schemes of cooperation among sellers.

In the Hotelling location game, prices are fixed and the firms choose the most appropriate locations for their activities. Consumers or customers are distributed along the interval $[0,1]$ with a uniform density equal to one. The prices equal one, and production costs are zero. The players in the Hotelling game are sellers, say a , who simultaneously choose locations $x(a) \in [0,1]$. They are ordered by their location $x(1) < x(2) < \dots < x(n)$, $x(0) = 0$ and $x(n + 1) = 1$. Seller number, a , attracts half of the customers on the gaps on each side of him so that his payoff is $(x(a) - x(a - 1))/2 + (x(a + 1) - x(a))/2$.

The existence of an equilibrium in the Hotelling location game was investigated for homogeneous and discriminate pricing, elastic and inelastic demand, independent and interacting products and bundles of products, non-differentiated and differentiated consumers, various distributions of consumers, different models of distance, and different numbers of sellers (see Gabszewich et al. 1986 for an overview of results). Eaton and Lipsey (1975) proved the existence of equilibrium in pure strategies with more than three sellers. Dasgupta and Maskin (1986a, b) and

Simon (1987) investigated equilibrium properties in mixed strategies for any number of sellers in a space of any dimension. The main results are that there exist equilibria for a wide variety of assumptions. The case of three sellers seems to be an exception. In that particular case, there was a game of musical chairs in which there was always an incentive for a seller to change location among a limited set of cells.

In this chapter, we investigate a number of related modeling frameworks to illustrate the complexities of the Hotelling location game. The different models can be related to the classical questions in location theory formulated by Palander (1935), Hoover (1948), and others. Essentially, the setup means that space is subdivided into discrete cells. In the framework of the Hotelling classical ice-cream vendor problem, we think of a beach with people assigned to predetermined chairs or sunshades. In the game theory framework, the problems we will primarily address are what will happen when new sellers arrive at the marketplace, i.e., at the linear beach:

- Will there be a single stable equilibrium, or will there exist cycles with sellers roaming the beach in search of ice-cream buyers?
- What patterns will emerge in a situation when some sellers are fixed in space and others are mobile?
- What patterns will emerge under different assumptions about the mechanisms of interaction among sellers?

There will be homogeneous sellers and buyers, and the products will not be differentiated. Generalizations are not pursued here as they would cloud the results of the behaviors we are trying to model. We illustrate how spatial patterns emerge from different assumptions with the help of a specifically developed model; see Roy, Snickars, and Zaitseva (2000) for a full description of the analyses. Our basic problem concerns the operation of a market with multiple sellers and customers where the behavior of the sellers (e.g., ice-cream vendors) is driven by the objective to maximize market share and the behavior of the customers (e.g., visitors of the beach) is to maximize accessibility to purchase the product for sale (i.e., ice creams).

Since the behavior of sellers and customers is not complementary, we will consider a total of six frameworks, each one intended to model a particular behavioral paradigm. There are two different market types and three optimization strategies. The first optimization strategy will take a seller perspective, specifically optimizing for the last seller to enter the market (sellers do not cooperate). The second takes the customer perspective thus optimizing the total benefit to all customers. For the first market type, customers are assumed to only use the most accessible (or closest) seller. We will call this a closed market. In the second case, customers will share their purchasing among all sellers in proportion to their relative accessibility. We will call this an open market. The relationships between customer and seller behavior are shown in Table 43.1.

In Model 1, sellers are locating to maximize their market share and do not cooperate in sharing the market. The market share for a new seller is determined by maximizing the number of customers who are closer to the new seller than to each of the other sellers. We assume that a new seller entering the market has only the

Table 43.1 The considered location models and market types

	Seller perspective Noncooperative game	Customer perspective Welfare maximization
Closed market		
Customers use nearest seller only	Model 1	Model 2
Open market		
Customers use accessible sellers	Model 3	Model 4

choice of location to maximize the market share; no other mechanisms are possible (e.g., price or product differentiation). In Model 2, sellers are locating to maximize the (total) accessibility to all customers, with each customer choosing to use the nearest seller only. From the customers' view, a seller which is closer offers additional benefit. The accessibility is estimated by a negative exponential distance decay function.

In Model 3, sellers locate to optimize their market share but do not cooperate in sharing the market. The market share is being determined by the accessibility of sellers to customers. It is assumed that sellers can attract a proportion of all customers, depending on the relative accessibility of sellers to customers. In Model 4, sellers locate to maximize the total benefit to all customers. Customers will share their purchasing with all sellers in proportion to their relative accessibility as measured by a negative exponential distance decay function.

In the context of cellular automata, we will consider our market to be defined over a grid of cells, each occupied by a seller or customer (or being empty). From some initial state, we will examine how the system evolves as more sellers are added to the system. There are thus two cases to consider. The first assumes that once located sellers do not relocate. This framework attempts to model urban systems in a development phase or systems in which some seller units are fixed and others mobile. The second case assumes that after the addition of a new seller, there is some time period during which the urban system adjusts toward an equilibrium state which is facilitated by allowing all sellers to change location to improve their relative payoffs.

While these modeling frameworks represent just a small sample of possible options and behavioral assumptions, they will permit us to see how the cellular automata handle the different situations. Our objective is to demonstrate that a cellular automata approach can produce results consistent with what we should expect from classical theory or intuitively from a behavioral analysis.

43.4 Basic Modeling Principles and Assumptions

As the basis for modeling, we take a spatial framework based on a regular grid of cells, (see also Roy et al. 2000). Each cell represents a unit of space which may contain some particular urban activity. The spatial arrangement of cells reflects the spatial organization of a seller-customer system. One might consider two basic spatial arrangements. One is a classical one-dimensional model where the customers are

located along a line and the other is a two-dimensional problem with a square array of customers. The analysis permits a range of cell types and the specification of several operational parameters (e.g., accessibility indices, competition factors, allocation sequences, clustering mechanisms, and distance metrics). Each cell will only be allocated to one seller, but a seller may be located in the same cell as a customer. The assumptions we will make for the purpose of comparability are as follows:

- Distances are Euclidean and measured from cell center to cell center.
- The negative exponential distance attenuation parameter is taken as 1.0.
- In the one dimension, sellers locate along the edge of a line of customers.
- In two dimensions, sellers may overlay customers within occupied cells.

We begin with some general definitions:

A is the set of sellers, a being any seller, and b a new seller; $a, b \in A$.

S is the set of customers, s being any customer, $s \in S$.

a, b , and s represent locations, $x(a)$, $x(b)$, and $x(s)$.

$d(s, a)$ represents the distance from a customer at cell s to a seller at cell a .

$W(a)$ represents the attractiveness of cell a for a new seller.

W is the attractiveness summed over all cells of the total system.

Model 1: Closed market, seller perspective and noncooperative game

Given that a system exists with a number of sellers and customers, an additional seller will locate at cell a when this cell maximizes the market share for the new seller. Sellers do not cooperate in any way. The market share is determined from the number of customers closer to k than any other seller. Hence a will be chosen as follows:

$$\begin{aligned} W &= \max W(a), \text{ across all } a \\ W(a) &= \sum \underline{d}(s, a), \text{ across all } s \\ \underline{d}(s, a) &= 1, \text{ if } d(s, a) < \min d(s, b), \text{ across all } b \text{ other than } a \\ \underline{d}(s, a) &= 0, \text{ otherwise} \end{aligned} \tag{43.1}$$

The new seller adopts a selfish view, attempting to claim as much of the market share as possible, knowing that if she locates so that the cell is closer to a customer than any other seller, then she will claim all the purchases of that customer. Naturally, the other sellers will not be content with the situation and, if possible, attempt to relocate to reclaim some of their lost market shares.

Model 2: Closed market, customer perspective and welfare maximization

This model takes the customer perspective. The location of a new seller is taken to maximize the accessibility of customers to sellers. Customers do not care which seller they use, but they will choose the closest and will (collectively) be more satisfied if the total accessibility is maximized. The attractiveness W is defined as follows (μ is a distance attenuation parameter):

$$\begin{aligned} W &= \max W(a), \text{ across all } a \\ W(a) &= \sum (\exp(-\mu \underline{d}(s, a))), \text{ across all } s \\ \underline{d}(s, a) &= \min d(s, b), \text{ across all } b \end{aligned} \tag{43.2}$$

This implies that the new seller will be located at cell a, which results in the total accessibility for all customers being maximized. Each customer chooses the closest seller exclusively for their purchases. Allowing existing sellers to relocate in response to this new seller entering the system may result in further improvements to the collective payoff to customers. Sellers placed closer to customers are considered more beneficial in accordance with the posited distance attenuation function.

Model 3: Open market, seller perspective and noncooperative game

Here we assume that customers are prepared to share their custom over all sellers in proportion to their relative accessibility to sellers. When a new seller locates (at cell a), he/she can count on capturing a proportion of all customers' business and so attempts to maximize this share. The proportion is assumed to be based on the relative accessibility of sellers to customers as computed from a standard distance-based accessibility measure. The new seller is thus located at cell a so that

$$\begin{aligned} W &= \max W(a), \text{ across all } a \\ W(a) &= \frac{\sum \exp(-\mu d(s, a))}{\sum \exp(-\mu d(s, b))}, \text{ across all } b \text{ and } s \end{aligned} \quad (43.3)$$

As with Model 1, the location of a new seller will most probably reduce the market share of the remaining sellers, who may then wish to relocate in an attempt to minimize this loss. In the accessibility case, the total purchases from the customers will depend on the number of sellers unlike in the closed market case when the total demand in the system will stay the same irrespective of the total number of sellers and their locations.

Model 4: Open market, customer perspective and welfare maximization

In this final model, the location of the new seller is taken to maximize the collective payoff to all customers assuming that customers will share their purchasing power with all sellers in proportion to the relative accessibility of sellers to customers. In this case, we have, therefore, the new seller being located at cell a so that

$$\max W(a) = \frac{\sum \exp(-\mu d(s, a))}{\sum \exp(-\mu d(s, b))}, \text{ across all } s \quad (43.4)$$

As with the previous models, if existing sellers are permitted to relocate, further improvements in total customer payoffs may be possible.

43.5 Experimenting with the Hotelling Model

To study the behavior of these models, an experimental test bed will be established. It may be, for instance, that the end result is path-dependent and thus influenced by the initial spatial distribution of the sellers. There could also be deviations between theory and practice because of our decision to use discrete cells rather than a continuum of possible locations (see also Puu 2003).

We are interested in comparing the cellular automata results with the classical theories as described earlier in the chapter. Our linear region must be of finite size to be computationally manageable. This fact will naturally introduce some edge effects due the size of the system. We expect these will not cause fundamental problems, providing the models are sufficiently large (i.e., have a large enough number of cells) relative to the number of sellers added to the system. The likely effect of a small-size sample in this case will be that occasionally payoffs will not vary across location cells. We will therefore expect there to emerge several equilibrium patterns of location. This is also confirmed from performing theoretical experiments with the models in the framework of game theory. In these experiments which have been done for the one-dimensional case, the result is that the best-response functions of the sellers will contain sets of locations with equal payoffs; see the theoretical considerations in, for instance, Rasmusen (1989).

A best-response function reveals the best response of one seller given the locations of all other sellers. Since these best responses are set valued, it is to be expected that the simulations will indicate the existence of several possible equilibrium situations or situations in which sellers cycle between different locations. An analysis has been performed to compute the equilibrium state in mixed strategies for some simple cases of the model. They show that already in the case of two sellers, the classical Hotelling solution will not always appear simply as a result of the existence of several best-response locations under the metric used.

The linear model consists of a line of 20 customer cells, with (initially) one seller located at the fourth cell from the left (see the sequence of figures below). Three more sellers are then allocated to the system. The displays show the location of sellers, assuming all other cells are housing customers. We consider two cases, one where sellers are fixed and cannot relocate once making an initial decision and one where they can and, indeed, generally do relocate. As a sensitivity test, simulations are performed also for other initial positions of the first seller.

In the first case, a single seller is placed in the fourth slot, and the positions for the second and subsequent allocations are computed for the given system state. In the second case, the initial position for the new seller is computed and the seller allocated. Then the positions of each seller are reviewed (in turn), and the sellers are relocated if better positions can be found. This relocation process is repeated iteratively until an equilibrium is obtained (i.e., no locational changes for any sellers can improve his/her individual payoffs). In some cases, a unique equilibrium is not obtained as the allocation pattern cycles through a sequence of cells.

In Figs. 43.4–43.7, the seller positions are shown shaded. Where a final stable equilibrium state is not found, the sellers tend to cycle through a number of states, the range of which is shown by the more lightly shaded cells. The darker shaded cells show a typical (but no equilibrium) state. The lack of convergence is not unexpected as we are dealing with a system with discrete spatial positions and a relatively limited number of customers compared to sellers.

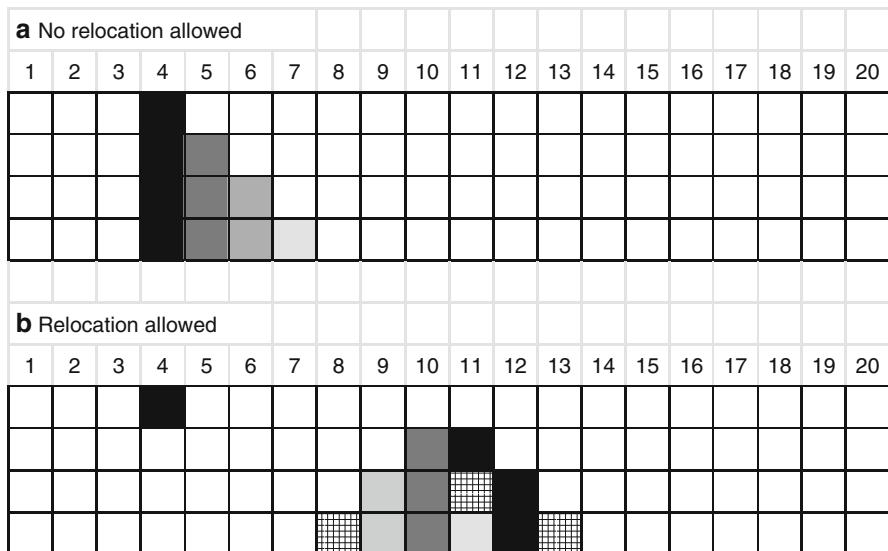


Fig. 43.4 Closed market, seller perspective and noncooperative game (Model 1): (a) sellers fixed after initial allocation and (b) sellers reallocated after new seller entered. The vertical dimension represents end situation after additional sellers have entered. The hashed areas represent cells in which cycles occur

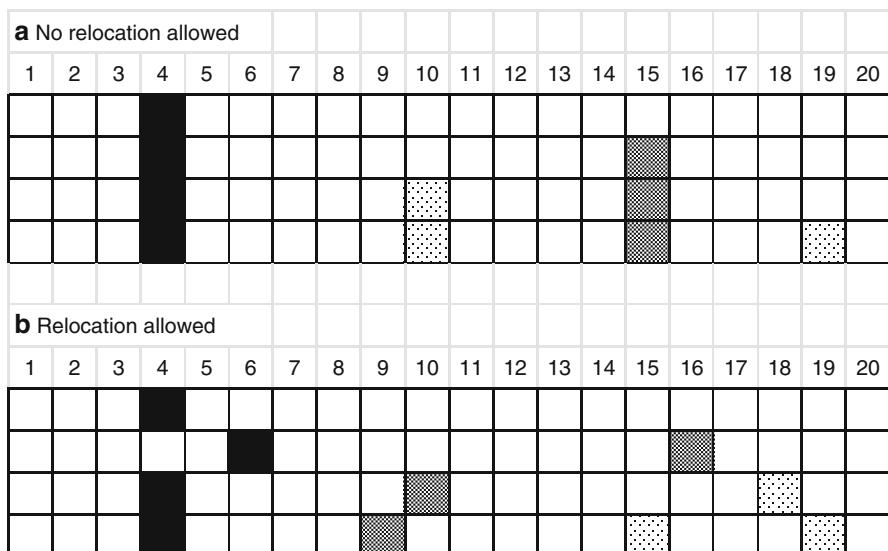


Fig. 43.5 Closed market, customer perspective and welfare maximization (Model 2): (a) sellers fixed after initial allocation and (b) sellers reallocated after new seller located. The vertical dimension represents end situation after additional sellers have entered

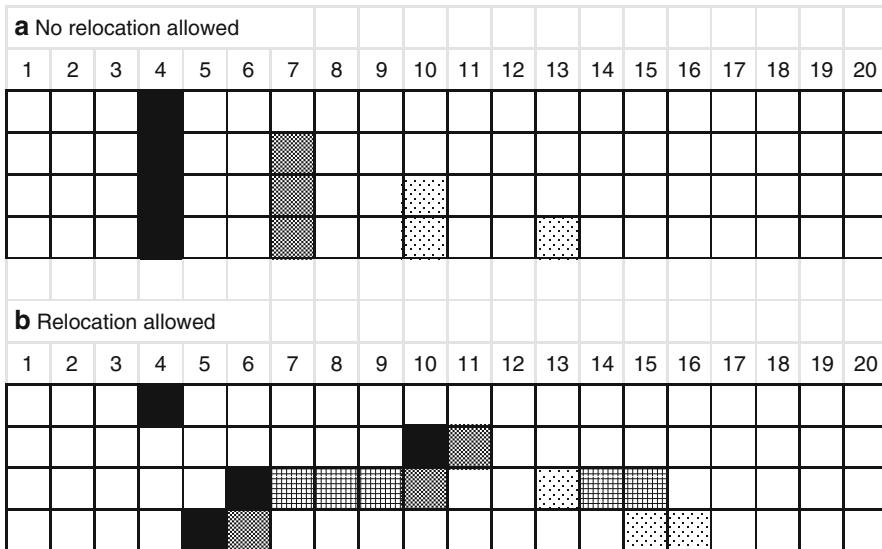


Fig. 43.6 Open market, seller perspective and noncooperative game (Model 3): (a) sellers fixed after initial allocation and (b) sellers reallocated after new seller located. The vertical dimension represents end situation after additional sellers have entered. The hashed areas represent cells in which cycles occur

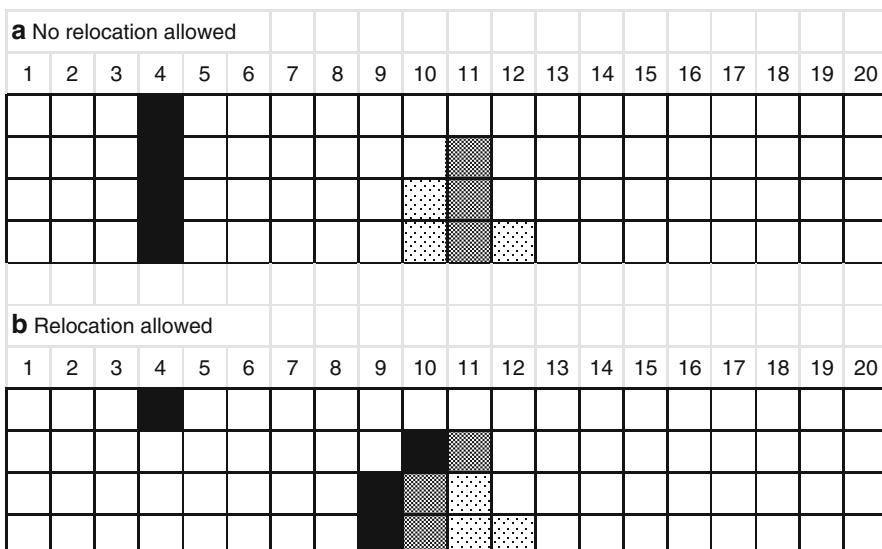


Fig. 43.7 Open market, customer perspective and welfare maximization (Model 4): (a) sellers fixed after initial allocation and (b) sellers reallocated after new seller entered. The vertical dimension represents end situation after additional sellers have entered

43.6 Analysis of the Simulation Results

We will now discuss the results in relation to a set of figures showing the end result of a number of simulations. The figures are organized so that each row represents the end state of the system with one, two, three, and four sellers. Note also that the market is represented by a one-dimensional array of 20 cells, each one of which can house one seller at a time.

Model 1 represents a selfish strategy for each new seller, implying that sellers do not cooperate to share the market. This can be seen in Fig. 43.4a. Each new seller takes a position immediately to the right of the previously allocated seller, thus claiming the market for all customers to her right. The other sellers naturally lose market share, and when allowed to relocate, as shown in Fig. 43.4b, they cluster at the center of the line. The second row shows the classical Hotelling result. The states for three or more sellers do not build up to a stable equilibrium since one, or more, of the sellers will always see a way of improving their market share (there are no relocation costs in our model). This is in line with the theoretical developments offered by, e.g., d'Aspremont et al. (1979).

It may be noted that the cycling in the fourth row will keep the four sellers as close to one another as possible in the middle of the market. The simulation shows, however, that the outermost spatial position will be challenged leading to some likelihood that this cell will also be occupied.

Model 2 takes the customer perspective, and thus we would expect to see the distribution of sellers optimized to suit the customers. This result can be seen in Fig. 43.5. When sellers do not relocate, Fig. 43.5a, the new seller locates in a way to split the longest line of sellers in half (approximately, of course). This is more clearly seen in Fig. 43.5b when sellers are allowed to relocate. This model gives results which seem to be directly following from the Hotelling problem. The welfare maximizing spatial pattern is such that the sellers cover the market rather than crowding toward the center of the joint market. The ultimate spacing is, of course, influenced by the fact that the market covers exactly 20 cells. It is to be noted that no instabilities occur in this model.

Model 3 takes a seller perspective, but this time opening the market so that we assume that customers will share their purchasing power among all sellers, but in accordance to their relative accessibility to each. The sellers do not cooperate to share the market. The result is shown in Fig. 43.6. The results are similar to the closed market case (Model 1) but with the sellers being more spatially distributed. This is to be expected as the sellers share the customers' market, making the choice of location less sensitive to claiming customers from existing sellers. The sellers can take customers from each other without piggybacking each other at the middle of the market.

Model 4 takes the customer perspective with the open market strategy. In the one-dimensional market case shown in Fig. 43.7, the results are quite clear. Sellers locate to maximize the total accessibility to customers. Since, from the customers view, the sellers are not competing with each other, the sellers tend to congregate toward the center of the linear system space. If one compares the end result for the

closed market case with the open market one, one can observe that the spatial patterns seem to be more concentrated out in the open market case than in the closed market one where sellers one access the nearest customers. The result will be sensitive to the choice of distance decay parameter to illustrate the attenuation of demand.

The simulation results above have been developed for the case of a linear market. Let us assume that the spatial competition takes place on a square with 10 cells in each direction and that the first seller is placed in cell (1,1). The end point for the noncooperative case in which sellers cannot relocate will be that late-coming sellers will take over the market by placing themselves on the diagonal from that corner. If the sellers are allowed to be relocated, they will cycle around the central cell (5,5). Thus, they will not be placed only in the middle area made up by the most central cells. One reason for this is the fact that the total market is limited in all directions.

In the case of welfare maximization, again starting with a seller in cell (1,1), the end result in the relocation case will be as expected that each seller will create a local monopoly at each of the four corners. In the case of fixed locations, the result is similar although influenced by the starting location of the first seller. The end results in the open market case are spatially more complex with cycles occurring.

A general conclusion from the experiments is that in simple cases, the theoretical results will be replicated. However, when complexity increases, the classical theories will lose some of their predictive power. The end results will be path-dependent, i.e., be different depending on the starting position of the first seller. The resulting patterns of spatial competition are not always stable, but sellers seem to cycle between a limited set of cells. This indicates the complexity of spatial competition and makes it necessary to develop more complex urban simulation models to attain further predictive power.

43.7 Conclusions

The modeling process by which the above results have been generated is quite simple. At the same time, the results are closely related to the classical problems of location theory posed by Hotelling, Hoover, and Palander. What we have illustrated is that we can replicate these results with good precision and efficiency using modern computational techniques. We have also demonstrated that modern theoretical work can get a substantial support for combining strict mathematical modeling with computer simulations. Finally, we have demonstrated the capacity to add complexity stepwise to show how further introduction of behavioral realism will affect the results both in terms of end-point equilibria and in terms of development paths in spatially competitive settings. It is also possible to use these methods to illustrate in a pedagogical way how different parameter settings will change the resulting spatial patterns.

The process of integrating different modeling frameworks into the cellular automata context is straightforward. We need to define the appropriate accessibility

function and the proper selection criteria for location choice of subsequent sellers. This function is applied to all cells (or at least to a sufficiently large number) to cover the region of interest. The state of the cell meeting the selection criteria is then designated to be a seller location. While the cellular automata framework facilitates a whole range of more complex analyses that might affect the choice of location, we have not included these here. For example, it is straightforward to introduce issues of price competition, cost of relocation, policy restrictions, clustering, and other nonmarket factors (e.g., collusion among a subset of sellers). The recent literature abounds with examples of such attempts as showed, for instance, by the reviews of Wegener (2004) and Batty (2008).

As a result of our modeling framework, we gain an immediate impression of the emerging spatial and temporal organization of the spatial competition system. We can observe the changes as they are computed and quite readily see if the behavior is in line with what might be intuitively expected. As a result, we can create an opportunity to explore a range of initial states and to evaluate the sensitivities of various modeling parameters on the simulation outcomes. Such a capability is a useful addition to available tools for the analysis of the dynamics of complex urban and regional systems.

It appears from our studies that cellular automata can produce results which seem plausible and demonstrate behaviors that we might have expected under classical theoretical assumptions. They also replicate the theoretical results in cases where strict comparisons can be made. This might be of value in studying the dynamics of complex processes and learning more about how they impact on the form and composition of urban regions. The cellular automata are conceptual tool that can be used in teaching or in visualizing behavioral dynamics. In this mode of use, the method will accompany theoretical analyses of urban systems, strengthening the theoretical insight about the behavior of urban agents.

The computations require information about the temporal state of cells and the ability to compute properties about neighboring cells. Such computations, as well as the graphical display of the cells with their properties, are generally well handled in geographic information systems which contain the spatial information. A limiting factor to treat different dynamic processes interdependently will be the computational speed available. Accessibility computations like those used here are very sensitive to the number of active cells in the system. Implementations will thus require some care to optimize computational processes even with modern computing speeds.

References

- Anas A (1987) Modelling in urban and regional economics. Harwood Academic, Chur
Batty M (2008) Fifty years of urban modelling: macro statics to micro dynamics. In: Albeverio S, Andrey D, Giordano P, Varcher A (eds) The dynamics of complex urban systems: an interdisciplinary approach. Physica, Heidelberg, pp 1–20
d'Aspremont C, Gabszewicz J, Thisse J (1979) On hotelling's stability of competition. *Econometrica* 47(5):1145–1150

- Dasgupta P, Maskin E (1986a) The existence of equilibrium in discontinuous economic games I: theory. *Rev Econ Stud* 51(1):1–27
- Dasgupta P, Maskin E (1986b) The existence of equilibrium in discontinuous economic games II: applications. *Rev Econ Stud* 51(1):27–41
- Downs A (1957) An economic theory of democracy. Harper and Row, New York
- Eaton BC, Lipsey R (1975) The principles of minimum differentiation reconsidered: some new developments in the theory of spatial competition. *Rev Econ Stud* 42(1):27–49
- Fujita M, Krugman P, Venables A (1999) The spatial economy: cities, religions and international trade. MIT Press, Cambridge, Massachusetts
- Gabszewicz J, Thisse J, Fujita M, Schweizer U (1986) Location theory. Harwood, New York
- Heffley D (1972) The quadratic assignment problem: a note. *Econometrica* 40(6):1155–1162
- Hoover EM (1938) Location theory and the shoe and leather industry. Harvard University Press, Cambridge
- Hoover EM (1948) The location of economic activity. McGraw Hill, New York
- Hotelling H (1929) Stability of competition. *Econ J* 39(153):41–57
- Koopmans TC, Beckmann M (1957) Assignment problems and the location of economic activities. *Econometrica* 25(1):53–76
- Leydersdorff L (2002) The complex dynamics of technological innovation: a comparison of models using cellular automata. *Syst Res Behav Sci* 19(6):563–575
- Liu Y (2009) Modelling urban development with geographical information systems and cellular automata. Taylor and Francis, Boca Raton
- Nickel E, Puerto J (2005) Location Theory: A unified Approach. Springer, Berlin and Heidelberg
- Palander T (1935) Beiträge zur standortstheorie. Almqvist & Wiksell, Uppsala
- Puu T (2003) Mathematical location and land use theory: an introduction. Springer, Berlin/Heidelberg/New York
- Rasmusen E (1989) Games and information: an introduction to game theory. Blackwell, Oxford
- Roy GG, Snickars F (1996) City life: a study of cellular automata in urban dynamics. In: Fischer M, Scholten H, Unwin D (eds) Spatial analytical perspectives on GIS. Unwin and Hyman, London, pp 213–228
- Roy GG, Snickars F (1998) An interactive computer system for land-use transport interaction. In: Lundqvist L, Mattsson L-G, Kim TJ (eds) Network infrastructure and the urban environment: advances in spatial systems modelling. Springer, Berlin/Heidelberg/New York, pp 350–370
- Roy GG, Snickars F, Zaitseva G (2000) Simulation modelling of location choices in urban systems. In: Fotheringham AS, Wegener M (eds) Spatial models and GIS: new potentials and new models. Taylor and Francis, London, pp 185–201
- Sembolini F (2000) The growth of an urban cluster into a dynamic self-modifying spatial pattern. *Environ Plan B Plan Design* 27(4):549–564
- Simon L (1987) Games with discontinuous payoffs. *Rev Econ Stud* 54(4):569–598
- Snickars F (1978) Convexity and duality properties of a quadratic intraregional location model. *Reg Sci Urban Econ* 7(4):5–19
- Thisse JF, Button K, Nijkamp P (1996) Modern classics in Regional Science: Location Theory. Edward Elgar, London
- Wegener M (2004) Overview of land-use transport models. In: Hensher DA, Button K (eds) Transport geography and spatial systems, Handbook 5 of the handbook in transport. Pergamon/Elsevier Science, Kidlington, pp 127–146
- White R, Engelen G (1993) Cellular automata and fractal urban form: a cellular modelling approach to the evolution of urban land-use patterns. *Environ Plan* 25(8):1175–1199
- White R, Engelen G (1997) Cellular automata as the basis of integrated dynamic regional modelling. *Environ Plan B Plan Design* 24(2):235–246

Johannes Bröcker

Contents

44.1	Introduction	852
44.2	Basics	853
44.3	Labor with Different Skills	857
44.4	Inefficient Migration	860
44.5	Two-Way Migration	860
44.6	Dynamics of Factor Mobility	864
44.7	Migration and Agglomeration	869
44.8	Conclusions	871
	References	872

Abstract

This chapter introduces into the theory of labor and capital movements between regions or countries. Movements of other mobile factors, in particular knowledge, are not dealt. After an introduction defining terms, it explains the basic factor mobility model assuming perfect competition and full factor price flexibility. Particular emphasis is given to the welfare results: Who are the winners and losers if factors are allowed to move and under what conditions does free mobility increase overall efficiency? We show how factor allocations deviate from an efficient outcome if the markets do not work perfectly. After studying factor mobility in a static framework, we extend the analysis to a dynamic framework. It is needed because investment decisions are forward looking. Investors compare present expenditures with present values of future returns. The same holds true for migration because migrants invest into human capital when they expend migration cost today in order to

J. Bröcker

Institute of Regional Research, Department of Economics, University of Kiel, Kiel, Germany
e-mail: broecker@economics.uni-kiel.de

earn a higher income in the future. In a final section, we study the role of factor mobility in New Economic Geography. A concluding section points to further topics not dealt with in this chapter.

44.1 Introduction

Factor mobility means that factors of production move across geographical space in the course of time. Labor, capital, and knowledge are the factors that are mobile, at least in principle. There are numerous obstacles to mobility that may go that far that factor mobility is completely prevented. If the factor is labor, its movement across space is called labor migration. Migration is a wider notion covering the change of peoples' place of residence, be they workers, nonworking family members, or other persons not in the labor force like students or pensioners (for a review of migration research, see Greenwood (2007)). Modern democratic societies grant citizens the right to freely choose the place of residence within the country or to leave the country, but restrict the right to freely enter the country and may restrict the free choice of residential location of persons who are not citizens of the country. In the European Union free mobility is, after a transition period, extended to the entire area of the union.

People either freely choose to change residential location because they expect better living conditions in the destination regions than where they hitherto have lived or they are violently forced. In the latter case they are classified as refugees or displaced persons and typically not called migrants, but there is no sharp borderline between migrants and refugees. About ten million people in the world are officially counted as refugees by the United Nations (UNHCR 2010). Regarding movement of persons, this chapter is confined to labor migration. We do not deal with migration of persons not in the labor force and exclude issues like displacement or fleeing from war or terror.

Capital mobility has aspects in common with labor mobility, but there are also important conceptual differences. While workers physically relocate, this is only exceptionally the case for capital. Reparations after the war are such an exceptional example. Though physical relocation is an exception rather than a rule, we nevertheless treat capital mobility in the simplest models as if capital was physically relocated. When treating capital mobility in a dynamic framework, however, we distinguish between real and financial capital. Financial capital is highly mobile in most parts of today's world, but real capital like buildings, fabrics, stocks of goods, and goods in process only relocates steadily through higher investment in one place and lower investment – including net disinvestment due to depreciation – in another place. In other words, real capital does not literally relocate, but its spatial distribution changes due to differential real capital growth.

An important subcategory of investment is foreign direct investment (FDI) meaning that some agent from country i (typically a firm) invests into another country j . This is to be distinguished from a capital flow in the way that an agent in country j invests, obtaining the financial means through the financial capital market, for example, by issuing bonds or selling shares of an incorporation. In the latter the

creditor has no role other than providing the financial means, while in the former the investment also benefits from nonfinancial resources of the investor like management capabilities, knowledge about products, technologies, and markets. The literature on motives and consequences of FDI has exploded during the last decades and is too wide to be dealt with here.

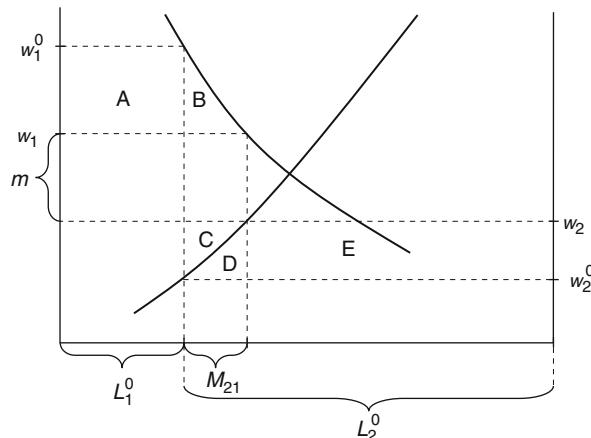
The third mobile factor mentioned above is knowledge. Partly it flows between regions carried in the brain of people, but knowledge flows have many other forms like learning and copying, trade in patents and blueprints, and trade in research and development services. Knowledge as a production factor is very different from other factors and can therefore not be dealt with by the same theories. Different from labor or capital, it is a non-rival input meaning that its use by one agent does not hinder its use by others. Furthermore, it is often difficult to privatize returns from investment into knowledge; others can copy without paying. Finally, investment into knowledge is usually much more risky than investing into other factors. An appropriate treatment of knowledge mobility would thus lead us into completely different realms of theory, in particular into endogenous growth theory, which is beyond the coverage of this chapter (see Acemoglu (2009) for a modern treatment).

44.2 Basics

Figure 44.1 illustrates the basics in factor mobility theory (see Borjas (1994, 2008) for an introduction). Consider a world of two regions, 1 and 2, both producing a single homogenous output freely tradable across regions. It is taken as the numéraire; its price is one. All factors but one, labor say, are immobile. Competition is perfect. Factor prices thus equal their respective marginal products. Let initially the total world stock of labor be distributed among the two regions such that the wage in region 1, w_1^0 , minus migration cost m exceeds the wage in region 2. In the figure, L_1^0 and L_2^0 denote initial factor endowments of regions 1 and 2, respectively. The width of the figure is the total stock of labor. Labor in region 1 (2) is measured from the left (right). The two curves are the marginal productivities of the two regions. They are at the same time the labor demand curves, falling with increasing labor employed in the respective region.

If labor is allowed to migrate, M_{21} workers move from 2 to 1 until the wage in region 1 net of migration cost just equals the wage in region 2, $w_1 - m = w_2$. Who will gain, who will lose? First, migrants obviously win $C + D = (w_2 - w_1^0)M_{21}$; otherwise they would not move. Workers staying behind gain the same per worker, that is, E in total; otherwise they would also move or, if they won more than the migrants, migrants would have stayed. Owners of immobile factors in region 2 lose E which is the income share shifted from other factor owners to the workers staying home. But they lose more, also D , because migrants contributed more than what they got paid when they worked in region 2. A worker got paid the contribution of the marginal worker, but intramarginal workers contributed more than the marginal worker due to decreasing marginal productivity. Taking the region of origin (called the “origin” in the following, for short) as a whole (workers plus other factor owners), it suffers the net loss D .

Fig. 44.1 Basics of factor migration theory: Given initial stocks of labor L_1^0 and L_2^0 in regions 1 and 2, respectively, the wage w_1^0 in region 1, net of migration cost m , exceeds the wage w_2^0 in region 2. After M_{21} workers moved from region 2 to region 1, the wage rate in region 1, net of migration cost, $w_1 - m$, equals the wage w_2 in region 2. The destination region 1 gains B , the origin region 2 loses D , and migrants gain C plus D



For the destination region (called the “destination” in the following, for short), the story is similar, with signs reversed. Workers lose A and other factor owners win A + B. Thus, the destination’s net gain is B. Summing up, the world society gains B + C :

$$\begin{array}{ll}
 B + C = -A & \text{loss of workers in 1} \\
 + A + B & \text{gain of other factor owners in 1} \\
 + E & \text{gain of workers in 2} \\
 - (E + D) & \text{loss of other factor owners in 2} \\
 + C + D & \text{gain of migrants.}
 \end{array}$$

$B + C$ is the integral of marginal productivity gains, net of migration cost, over migrants.

Workers staying behind and workers already residing in the destination before are the factors competing with the migrants; the other factors are jointly complementary to migrants. We can thus summarize what we found as follows: *If factors move, competing factors in the destination lose and in the origin win, complementary factors in the destination win and in the origin lose, destinations as a whole win, origins as a whole lose, migrants win, and the world society as a whole wins.* While this result is derived for labor mobility, it also holds for capital if capital rather than labor is the mobile factor.

Two assumptions are vital for this result: (i) factor demand curves are falling, and (ii) factor price differentials are the only migration motive. Assumption (i) makes sure that factor flows are self-defeating: The more workers or capital move, the lesser becomes the incentive to move. Agglomeration theory tells that this needs not be so if the basic neoclassical assumptions (constant returns to scale, perfect competition) are given up. We return to this issue in Sect. 44.7.

Assumption (ii) is obviously extreme, in particular as far as labor migration is concerned. For labor migration we therefore now take a second look at this assumption. Empirical research on migration uncovers many motives beyond (expected) income differentials. Still, there is unanimity among reviewers of empirical migration research that, after controlling for other variables, income differentials are among the most relevant migration incentives; possibly they are the one most relevant migration incentive. But there is also ample evidence of persisting wage differentials between regions, controlling for skill differentials, despite free labor mobility. Several reasons account for this observation:

- (a) *Amenities*: Workers are willing to accept lower-paid jobs if they get compensated by favorable living conditions such as provision of public goods, nature, and safety. This is easily incorporated into the above approach by shifting the labor demand curves upward or downward such that w does not represent just wage, but a wage corrected for the willingness to pay for amenities. All welfare results go through as stated.
- (b) *Consumer price differentials*, in particular land price differentials: These are also easily incorporated by correcting labor demand for consumer prices such that real instead of nominal wages appear on the vertical axis. Note, however, that, while workers care about real wages, the demand decision of firms depends on nominal wages. Land prices are endogenous. They are increasing in the number of workers entering the region. The real wage curve is thus steeper than the nominal wage curve. If scarcity of residential land is taken into account, the welfare gain in the destination does not only end up in the pockets of complementary production factors owners but also in those of residential landowners. Similarly, the welfare loss in the origin is partly passed on to the residential landowners. But apart from this modification, welfare results remain intact.
- (c) *Unemployment*: If labor markets do not clear, potential migrants weigh the income they could earn on a job in a destination with the probability of obtaining or keeping it. There is a wide literature claiming this to be the main explanation of persistent big wage differentials between rural and urban regions in less developed countries (Harris-Todaro hypothesis (Harris and Todaro 1970)). These countries are typically characterized by a dual labor market. In rural areas people are either self-employed or low-paid farm hands, and the wage is downward flexible such that there is little or no unemployment. In the cities there is a formal sector paying comparatively high wages that are not (fully) downward flexible. There is also an informal urban sector with low pay. Downward wage rigidity in the formal urban sector can have different reasons, minimum wage laws, state-owned firms with regulated wages, union power, or efficiency wages. The latter means that firms have an incentive to pay a higher-than-market-clearing wage in order to maintain the threat of a job loss as a disciplinary device to prevent workers from shirking or to force them to work harder. In this situation the migration equilibrium is different from what we have seen so far. Workers leave the rural region as long as the *expected* income y_j^e in city j ,

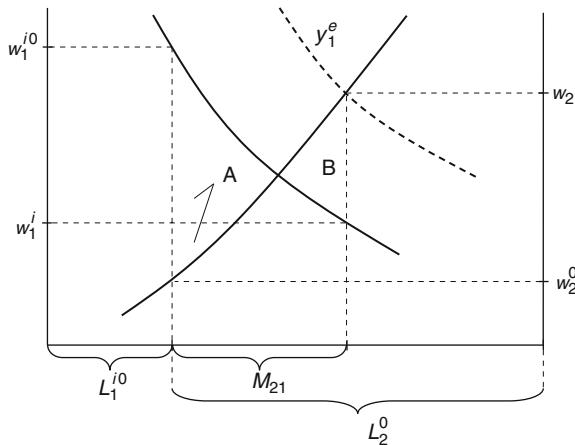


Fig. 44.2 Rural–urban migration in the Harris-Todaro model: Given initial stocks of labor L_1^{i0} in the informal urban sector and L_2^0 in the urban region, respectively, the expected urban income y_1^e exceeds the rural wage w_2^0 . After M_{21} workers moved from region 2 to region 1, both coincide. In addition to the wage w_1^i in the informal sector, the expected urban income also takes the chance to get a higher-paid job in the formal urban sector into account. There can be a total welfare loss if B is greater than A

$$y_j^e = p_j w_j + (1 - p_j) y_j^u$$

exceeds the rural income. w_j denotes wage in the formal sector; y_j^u is income of an unemployed (wage w_j^i in the informal sector plus transfer income, if there is any). p_j is the probability of having a job in the formal sector. It is decreasing in the rate of unemployment u_j . In particular, $p_j = (1 - u_j)$ if workers are randomly assigned to jobs.

Regarding the welfare impact, the gains and losses of owners of complementary and competing factors in the origin and destination have the same signs as before. But the country as a whole may lose rather than win, as illustrated in Fig. 44.2. Region 1 is the urban, region 2 the rural region. For the urban region only the informal labor market is shown with employment L_1^i and wage rate w_1^i . No changes need to be taken into account for the formal sector because wages and employment in that sector are unaffected by assumption. What is affected, however, is who happens to belong to the formal sector's labor force; migrants (though not all of them) enter it, but at the same time the same number of residents leave it. This affects individual but not total welfare.

For the sake of simplicity, Figure 44.2 is drawn for the case of zero migration cost. Rural–urban migration continues until the expected urban income y_1^e equals the rural wage w_2 . Thus, in equilibrium the urban wage in the informal sector is $w_1^i < w_2$. In the figure, w_1^i is larger than w_2^0 . But this is not necessarily the case. It may be smaller such that, without transfers, migrants not getting a job in the formal urban sector are actually worse off than if they had stayed home.

Still, they move because they expect having a chance to get a job in the formal sector, or because social benefits make them better off, or a combination of both. The total net welfare gain is $A - B$, which may be positive or negative. In any case the level of rural–urban migration is inefficiently excessive. The higher the urban wage in the formal sector and the larger the transfer income paid to an urban unemployed, the more the urban–rural migration exceeds its efficient level. Obviously, full wage flexibility in the formal sector is the first best solution of the problem. But this solution is typically not available, and countries therefore try to restrict rural–urban migration by other means such as China’s Hukou (household registration) system.

- (d) *Learning:* Jobs may offer non-visible benefits to workers letting them accept a lower real wage in the destination. Most important is the chance to learn on the job. It explains why people with higher education degrees but little experience on the job accept low real incomes in dynamic cities. They regard the income foregone a worthwhile human capital investment (Glaeser 1999).

44.3 Labor with Different Skills

The conclusion regarding the impact of immigration on workers’ welfare obtained so far is however too pessimistic. Immigrants are typically not perfect substitutes of residents. Often they are less educated or have acquired professional skills that the destination region is short of. Immigrants can be substitutes for residents, though not perfect substitutes. Depending on the degree of substitutability between labor and capital and between different kinds of labor as well as on input costs shares, wages of residents may either rise or fall as a response to immigration. To disentangle the parameter profiles leading to either an increase or decrease of residential wages, we extend the previous analysis to one with three factors, capital K , residential labor L_r (“residents,” for short), and labor L_i of immigrants (“immigrants,” for short). The respective wages are denoted w_r and w_i . The stocks of capital and residents are fixed, markets are perfectly competitive, and prices are perfectly flexible. The three factors are used for producing a single homogenous output taken as the numéraire (see Borjas (2003) for a more general analysis along these lines).

We assume a nested production function as illustrated in Fig. 44.3. Each node represents a quantity (output x , composite labor L , and the two kinds of labor, residents L_r and immigrants L_i) with associated prices (output price p , composite labor price w , and wages w_r and w_i for the two kinds of labor, respectively). For producing the output quantity x , capital K and composite labor L are combined in the upper nest, and composite labor is in turn composed of the two kinds of labor in the lower nest. The output price (equal to one by convention) is the minimal expenditure for capital and composite labor needed per unit of output. The composite labor prize w is the minimal expenditure for the two kinds of labor needed per unit of composite labor.

Fig. 44.3 Nested production function: Output x , sold at price p , is produced with capital K , rented at rate r , and composite labor L hired at composite wage w . L is a composite of labor supplied by residents (L_r) and immigrants (L_i), hired at wages w_r and w_i , respectively

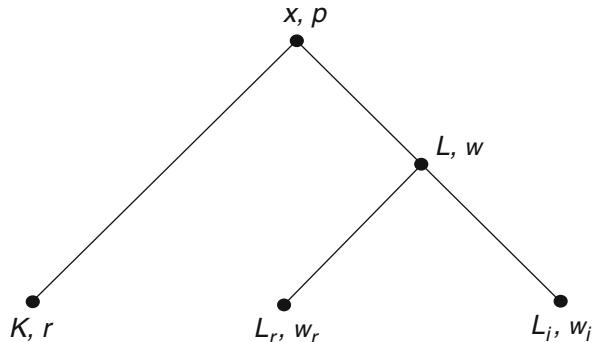


Table 44.1 Elasticities with respect to the stock of immigrants

	Variable	Elasticities
r	Rental rate	$S_i S_L / \sigma > 0$
w	Composite wage	$-S_i S_K / \sigma < 0$
w_r	Resident's wage	$S_i(1/\eta - S_K / \sigma) ?$
w_i	Immigrant's wage	$-S_i S_K / \sigma - S_r / \eta < 0$
x	Output	$S_i S_L > 0$
L	Composite labor	$S_i > 0$

[Table 43.1](#) shows how the endogenous variables respond to immigration. The entries in the table are elasticities of the respective endogenous variables with respect to the stock of immigrants. $S_K > 0$ and $S_L > 0$ are the shares of capital and labor cost in the cost of the final output, respectively ($S_K + S_L = 1$). Similarly, $S_r > 0$ and $S_i > 0$ are the shares of residential and immigrating labor in the cost of composite labor, respectively ($S_r + S_i = 1$). Furthermore, σ is the elasticity of substitution between capital and composite labor; η is the elasticity of substitution between the two kinds of labor. The former measures the percentage increase of the labor to capital ratio as a response to a one percent increase in the ratio of the rental rate to composite wage (similarly for the latter). For example, the first entry $S_i S_L / \sigma$ in the table is the elasticity of the rental rate with respect to the stock of immigrants, that is, the percentage response of the rental rate to a one percent increase in the stock of immigrants.

To derive these elasticities, we use the following facts on cost minimization of a competitive firm with output $x = f(z_1, z_2)$, inputs z_1 and z_2 , output price p , input prices w_1, w_2 , and linear-homogenous production function f :

- (i) $\hat{p} = S_1 \hat{w}_1 + S_2 \hat{w}_2$ with expenditure shares $S_1 \geq 0, S_2 \geq 0, S_1 + S_2 = 1$.
- (ii) $z_i = \hat{x} + \sigma(\hat{p} - \hat{w}_i)$, $i = 1, 2$, with elasticity of substitution σ .

$\hat{p} = dp/p = d\log(p)$ denotes the relative change of p (similarly for the other variables). Both facts are intuitive: (i) states that the percentage change of output price is the weighted average of the percentage changes of the input prizes, with

weights equal to the respective cost shares. (ii) states that the percentage change of the input equals the percentage change of the output corrected for the effect of the relative input price. The larger the elasticity of substitution, the stronger is the price effect. Rule (i) holds generally, for any functional form. While rule (ii) is typically derived assuming a CES (constant elasticity of substitution) form, it generally holds in the two-input case.

Writing the equilibrium as a total differential in logs (i.e., in relative changes) yields

$$\hat{p} = S_K \hat{r} + S_L \hat{w} \quad (44.1)$$

$$\hat{w} = S_r \hat{w}_r + S_i \hat{w}_i \quad (44.2)$$

$$\hat{K} = \hat{x} + \sigma(\hat{p} - \hat{r}) \quad (44.3)$$

$$\hat{L} = \hat{x} + \sigma(\hat{p} - \hat{w}) \quad (44.4)$$

$$\hat{L}_r = \hat{L} + \eta(\hat{w} - \hat{w}_r) \quad (44.5)$$

$$\hat{L}_i = \hat{L} + \eta(\hat{w} - \hat{w}_i) \quad (44.6)$$

Rule (i) is applied in Eqs. (44.1) and (44.2), rule (ii) in the other equations. Setting $\hat{p} = \hat{K} = \hat{L}_r = 0$, we solve for the entries in Table 44.1: \hat{r}/\hat{L}_i , \hat{w}/\hat{L}_i , and so forth.

All signs but one are unambiguous: The rental rate, output, and composite labor go up; the immigrants' wage and the composite wage go down with immigration. Quantity effects just depend on the respective input shares. The smaller the elasticities, the bigger are the price effects. Smaller elasticities require large price changes for the economy to adapt to changing input stocks.

Regarding the welfare effect for immigrants on the one hand and the collective of the other factor owners on the other hand, Figure 44.1 still applies. Immigrants compete with immigrants (their perfect substitutes). The more they are, the lower their wage. The collective of other factor owners gains because the average extra output generated by the immigrants exceeds the marginal output which is paid to them as their wage. Among the other factors, capital owners are sure to gain, but residents may gain or lose. They gain if and only if $\sigma/\eta - S_K > 0$. In particular, they always gain if $\sigma > \eta$, that is, if labor and capital are more substitutable than the two kinds of labor. They also gain if capital has a low share in factor costs.

All we have found so far holds, with all signs reversed, in the emigration region: Workers who are perfect substitutes of the emigrants gain, capital owners lose, and workers of a different type than those emigrating either gain or lose. The more complementary to the emigrants they are, the more they lose.

44.4 Inefficient Migration

We have seen that rural–urban migration is inefficiently large if the formal urban sector pays a downward nonflexible wage. But in a perfectly competitive world with flexible factor prices, free factor mobility brings about an efficient factor allocation across regions, though not everyone is made better off. Why do we not observe free mobility, in particular no free labor migration across national borders, and why do governments try to prevent big migration waves, for example, in Germany after unification or in the European Union after Eastern expansion? One obvious reason is the political economy induced by the distribution effects explained above. Workers competing with immigrants are many, while owners of complementary factors are few. In a democratic society the former are the majority of voters, opting for parties pursuing restrictive immigration policies.

Another reason is the aim to avoid excessive migration beyond the level that is welfare enhancing. Besides by unemployment, inefficiency of migration can be caused by external effects. If immigrants exert negative external effects as, for example, congestion of public infrastructure or social conflict in the destination, then restricting the number of immigrants is socially efficient. In addition, there may also be negative external effects of outmigration for the origin, as, for example, the loss of knowledge spillovers if educated workers leave the region. This strengthens the argument for restrictive migration policies.

This argument is however questionable for two reasons. First, there are also positive external effects for the destination. Immigrants may as taxpayers contribute more to the provision of public goods than is required to compensate residents for increased congestion. The more likely this is, the less public goods are subject to users' rivalry. Another positive externality in the immigration region is related to the brain drain. High-skilled workers are supposed to marginally contribute more than their respective gross wage to output because of knowledge spillovers, which favor residents without them having to pay for.

Second, there are also external effects in the region of origin. For example, while residents in the destination suffer from increased congestion (if not compensated by taxes of the immigrants), suggesting migration to be inefficiently large, the opposite happens in the origin, *per se* suggesting migration to be too small. It is therefore impossible to come to an *a priori* unambiguous conclusion as to whether and to what extent migration leads to inefficient labor allocation across regions or countries. An empirical assessment of external effects is needed for the concrete case.

44.5 Two-Way Migration

So far, migration seems to be a one-way road from a low-wage to a high-wage region. In some historical periods migration was in fact predominantly one way, from European countries to North America in the nineteenth century, from southern to northern Europe in the 1960s and early 1970s of the twentieth century or the East–West migration since the ending of the 1980s of the twentieth century. These

big waves fit well with the picture revealed by the model of the previous sections. They represent responses to severe regional or national disparities in expected lifetime incomes after removal of migration barriers.

Considerable migration flows are however also observed under conditions of moderate disparities and without any previous barriers having been lifted. Such flows are typically more balanced, going from one region to another as well as in the reverse direction.

The obvious explanation of this phenomenon is heterogeneity of workers regarding their respective preferences for different types of jobs. Consider an economy with n regions. Region i is the location of N_i firms with identical technologies. Each firm's labor demand is $l_i = Aw_i^{-\varepsilon}$. Initially, L_i^0 workers reside in region i . They migrate to the most attractive destination within the period considered. Workers have heterogeneous preferences regarding the attractiveness of jobs in the different firms. They are willing to accept lower payments, if they find a job more attractive. The attractiveness of a job in firm f compared to firm g , say, is quantified by the wage reduction a worker is willing to accept when choosing a job in firm f rather than firm g . For destination choice, migration cost is also taken into account. Both, migration costs and attractiveness are measured as a percentage of the destination wage. Let m_{ij} denote the share of migration costs for moving from i to j in region j 's wage rate. Furthermore, let e_{fh} be the share in the wage rate representing the attractiveness of a job in firm f for worker h . Thus, the wage plus attractiveness term minus migration cost is $(1 - m_{ij})w_j(1 + e_{fh}) \approx (1 - m_{ij})w_j \exp(e_{fh})$, if worker h initially resides in region i and the firm's location is j . One can show that, if e_{fh} is a Gumbel-distributed random variable, independently and identically distributed across all worker-job pairs, then we obtain the expected migration flow from i to j as

$$M_{ij} = L_i^0 \frac{N_j \mu_{ij} w_j^\lambda}{\sum_k N_k \mu_{ik} w_k^\lambda} \quad (44.7)$$

where $1/\lambda$ is the standard deviation of e times $\sqrt{6}/\pi$ and $\mu_{ij} = (1 - m_{ij})^\lambda < 1$. Note that i and j are allowed to be the same. Thus, M_{ii} denotes the number of workers residing initially in i and staying there. In equilibrium the labor market clears. Hence,

$$\sum_i M_{ij} = N_j A w_j^{-\varepsilon} \quad (44.8)$$

Equations (44.7) and (44.8) jointly determine the migration flows and the wage rates, given the parameters, initial number of workers, and number of firms in the regions. Another way to write the model is

$$M_{ij} = L_i^0 a_i \mu_{ij} N_j b_j \quad (44.9)$$

with constraints

$$\sum_j M_{ij} = L_i^0$$

and

$$\sum_i M_{ij} = AN_j b_j^{-\varepsilon/\lambda}$$

with $b_j = w_j^\lambda$.

This form shows the equilibrium to be described by a constrained gravity model. L_i^0 and N_j are the masses, μ_{ij} is the resistance term, and a_i and b_j are the balancing factors relating to the regions of origin and destination, respectively. The gravity model is constrained on both sides, inelastically so on the origin side and elastically constrained on the destination side.

The formal structure of this model turns out to be a special case of a general interaction model suggested by Alonso (1978), called “A Theory of Movements” after the title of Alonso’s original contribution. Alonso introduces exogenous characteristic variables of origins and destinations corresponding to our masses L_i^0 and N_j , endogenous multipliers associated with origins and destinations, respectively, and elasticity parameters controlling the response of flows to these multipliers. Our model emerges if the elasticities on the origin side are set equal to zero. Alonso’s multipliers associated with the destinations – paraphrased vaguely as “competition, crowding, congestion” variables in his original paper – turn out to be increasing transforms of the wages. Introducing wages and the labor market equilibrium provides a theoretical underpinning of Alonso’s interpretations of the multipliers.¹

Gravity models are the workhorses of econometric migration research. Taking logs in Eq. (44.9) and adding a random term lead to a linear regression. Adding a time dimension allows for applying panel methods. Origin and destination characteristics that are constant over time can be represented by fixed effects. Instead of taking logs one can stick to the nonlinear logit form. A well-specified logit model for bilateral migration flows clearly shows the positive impact of wages and the negative impact of unemployment on destination choice (Davies et al. 2001).

Two extreme cases of the equilibrium are of special interest: If (for finite λ) ε goes to infinity, the labor demand curve is horizontal. This implies a fixed b and no demand constraint. If ε is zero, the labor demand curve is vertical. The equilibrium

¹To be precise, Alonso’s multiplier c_j is $c_j = w_j^{\lambda+\varepsilon}$, and his elasticity β is $\beta = \lambda/(\lambda + \varepsilon)$. Hence $c_j^\beta = w_j^\lambda$ and $c_j^{\beta-1} = w_j^{-\varepsilon}$.

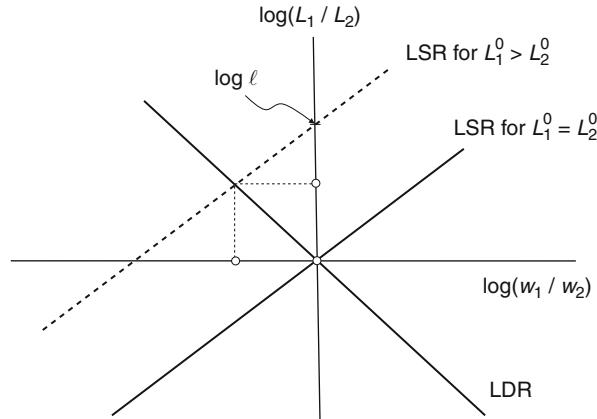


Fig. 44.4 Two-way migration: The equilibrium distribution of labor between the two regions (L_1/L_2) is attained where the *LSR* (labor supply ratio) and *LDR* (labor demand ratio) lines cut. Solid and dashed *LSR* lines refer to symmetrical and asymmetrical initial distributions of labor, respectively. *LSR* is increasing; *LDR* is decreasing in the wage ratio w_1/w_2 . In the symmetrical case, equal flows of migrants move both ways. In the asymmetrical case more workers move from 1 to 2 than the other way, but migration in one period does not fully equalize labor distribution and wages among the two regions

stock of labor in each region is thus fixed; the model is equivalent to the so-called doubly constrained gravity model in this case.

Let us work out the model for a world with two regions, covering the same number of firms each. Firms are identical, except for the fact that workers value jobs differently, as described above. Changing jobs within a region is costless, while migration to the respective other region incurs costs equal to m times the destination's wage. Then the number of jobs offered in region 1 relative to region 2 is

$$L_1^d/L_2^d = (w_1/w_2)^{-\varepsilon}$$

while labor supply in region 1 relative to region 2 is

$$\frac{L_1^s}{L_2^s} = \frac{M_{11} + M_{21}}{M_{12} + M_{22}} \quad (44.10)$$

with

$$M_{ij} = L_i^0 \frac{\mu_{ij} w_j^\lambda}{\sum_{k=1}^2 \mu_{ik} w_k^\lambda}, \quad \mu_{ij} = \begin{cases} 1 & \text{if } i = j \\ (1 - m)^\lambda < 1 & \text{else} \end{cases}$$

Figure 44.4 depicts the equilibrium. Axes are in log scale. The labor demand ratio (LDR) L_1^d/L_2^d is decreasing, and the labor supply ratio (LSR) L_1^s/L_2^s is increasing in

w_1/w_2 . Furthermore, as LDR goes to infinity (zero) and LSR goes to zero (infinity) if w_1/w_2 goes to zero (infinity), there is unique equilibrium. In particular, if $L_1^0 = L_2^0$ (solid LSR curve), then the equilibrium is the origin in the graph. It is symmetrical: $w_1 = w_2$, $L_1 = L_2$, $M_{11} = M_{22}$, $M_{12} = M_{21}$, and the out-migration rates are

$$\frac{M_{12}}{L_1^0} = \frac{M_{21}}{L_2^0} = \frac{\mu}{1 + \mu} \quad (44.11)$$

Clearly, the higher the migration cost and the more homogenous the job preferences, that is, the larger λ , the smaller are the outmigration rates.

The figure also shows an equilibrium for an initially unequal distribution of labor, $L_1^0 > L_2^0$ (dashed LSR curve). The LSR curve cuts the abscissa at $\log \ell$. ℓ denotes the LSR for $w_1 = w_2$:

$$\frac{L_1^0}{L_2^0} > \ell = \frac{L_1^0 + \mu L_2^0}{\mu L_1^0 + L_2^0} > 1$$

It follows that in the equilibrium, indicated by dotted lines, we find

$$L_1^0/L_2^0 > L_1/L_2 > 1 \text{ and } w_1 < w_2$$

Migration leads to a more equal, though not perfectly equal, distribution of labor. The distribution remains unchanged if m goes to one, of course, and it becomes equal if m goes to zero. The lower the migration cost is, the more migration tends to equalize the distribution of labor between the two regions. One can apply the model iteratively, period by period, such that the equilibrium in the first period becomes the initial distribution in the second and so forth. Then, period by period, the distribution equalizes more and more. In the limit we come back to the symmetrical equilibrium with migration rates as given by Eq. (44.11).

44.6 Dynamics of Factor Mobility

In the previous sections we studied factor movements as if the economy existed for only one period. In the beginning of the period, the distribution of factors across regions differs from what it would be if factor owners could freely decide where to locate their respective factors, labor of certain skills or capital. Then mobility restrictions are lifted or reduced, and factors jump to the place where they earn the highest real return net of migration cost. This kind of story is a shortcut to the long-run steady-state impact uncovered by a more sophisticated dynamic analysis.

Hence, let us extend the foregoing analysis to a dynamic framework (see Barro and Sala-i-Martin (1995, Sect. 9.1 and 3.5) for a textbook treatment of dynamic factor mobility). We begin with labor migration. To simplify, we abstract from land as consumption good and all kinds of market imperfections. Consider a small open region facing a fixed world wage rate \bar{w} . Capital is perfectly mobile such that the

interest rate in the region is the same as in the rest of the world, \bar{r} . It is constant from the region's point of view. Labor is mobile between the region and the rest of the world. At least one other factor (land) is immobile. Labor may be just a segment of the entire labor force.

Sjaastad (1962) introduced the idea that the decision to migrate is a decision to invest into human capital. Migrants compare migration cost with future returns. A potential migrant from the rest of the world enters the region at time t if the present value of the wage differential as of time t between the region and the rest of the world,

$$B(t) = \int_t^{\infty} (w(L(\tau)) - \bar{w}) \exp[\bar{r}(t - \tau)] d\tau \quad (44.12)$$

exceeds the migration cost $m(t)$. While we confine the discussion to the immigration process starting in a world where $w(t_0) > \bar{w}$, outmigration can be studied in a similar way if $w(t_0) < \bar{w}$ at start point t_0 . For the sake of simplicity, we assume the potential migrant to take wage differentials through the infinite future into consideration. One justification is that this reflects the fact that migrants not just care about their own but also their decedents' well-being. Another is that the approach implies that migrants move early in life, because the earlier they move, the longer they reap the benefits from a wage differential. For a young migrant the difference between an infinite and an end of active-life horizon is small.

If we assumed fixed migration cost m at all times, we would still observe migrants to jump in instantaneously and stay forever. The equilibrium labor force L^* would be obtained from the steady-state condition:

$$B = \frac{w(L^*) - \bar{w}}{\bar{r}} = m$$

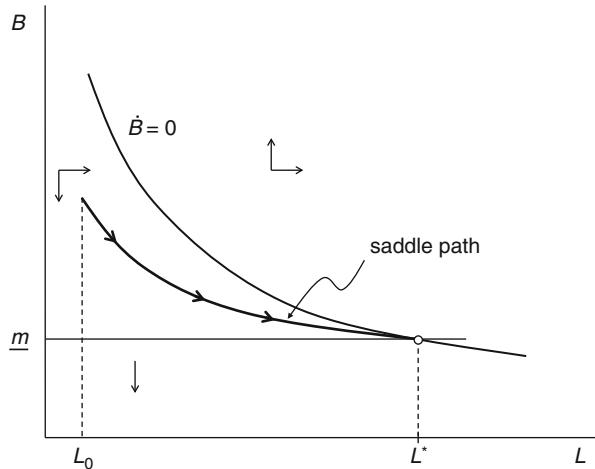
As this had to hold forever, the time argument is dropped. We may rewrite this as

$$w(L^*) = \bar{w} + \bar{r}m \quad (44.13)$$

bringing us back to the static-migration model in Sect. 44.2. The only difference is that migration costs would now be a per annum flow, namely, the perpetuity equivalent to the one-off cost m . The immigration would be $L^* - L_0$, where L_0 is the stock of labor at the start point t_0 .

Why do we observe a continuous migration flow rather than jumps in the regional stock of labor? One obvious reason is that new generations enter the pool of potential migrants only gradually. Still, also within a cohort reaching the age to look at the first job, not everyone moves at the same time. Some move as early as they are allowed to, others move later, and many do not move at all. Partly, the lack of information or the time needed to decide can explain this phenomenon. Another explanation is adjustment cost. If everyone expecting a gain moves at the

Fig. 44.5 Dynamic migration theory: In the course of time, the present value B of immigration benefits and labor L in the region move along the saddle path until B is down at the lower bound \underline{m} of migration cost, and immigration ceases. As indicated by the arrows pointing into the direction of movement, the saddle path is the only path leading to the long-run equilibrium



same moment, the region's labor market, housing market, etc. have to adapt instantaneously. This burdens migrants with costs. The more people move at the same time, the higher are these costs. It is thus reasonable to assume migration costs to be increasing in M/L :

$$m(t) = c(M(t)/L(t)) \quad (44.14)$$

$c(\cdot)$ is increasing and assumed to go to a lower bound $\underline{m} > 0$ if M/L goes to zero. Even if there is no other migrant, an infinitesimal migrant is facing strictly positive migration cost. She will stay if $B < \underline{m}$.

Noting that $M(t) = \dot{L}(t)$, where dotted variables denote time derivatives, Eqs. (44.12) and (44.14) can be written as a dynamic system in B and L :

$$\dot{B} = \bar{r}B - (w(L) - \bar{w}), \quad \dot{L} = Lc^{-1}(B)$$

The first equation is the time derivative of Eq. (44.12); the second is the solution of Eq. (44.14) for $M = \dot{L}$, using the equilibrium condition $m(t) = B(t)$ to substitute B for m .

Figure 44.5 illustrates the dynamics in a phase diagram. Arrows point into the direction of movement of B (vertical) and L (horizontal), respectively. Along the $(\dot{B} = 0)$ isocline, B does not change, that is, $B = (w(L) - \bar{w})/\bar{r}$. The isocline is decreasing because of decreasing marginal productivity. It cuts the $(B = \underline{m})$ line at the stationary state where condition (44.13) holds. In the dynamic transition B and L move along the saddle path asymptotically toward the steady state. If, starting at $L_0 = L^*$, regional productivity goes up due to an exogenous shock, then the $(\dot{B} = 0)$ isocline as well as the saddle path shift upward; B jumps up to the new saddle path and workers start moving into the region until the stock of labor has asymptotically adjusted to the new higher stationary state.

A similar theory, called Tobin's q-theory, explains why capital does not jump to a place of higher marginal returns but adjusts by a smooth growth of the capital stock in the destination. Consider a small open region with capital stock K and marginal capital productivity $f(K)$, $f(\cdot)$ positive and decreasing in K .² Financial capital is perfectly mobile. The uniform world interest rate is \bar{r} . The increase of the capital stock per unit of time \dot{K} is gross investment I minus depreciation δK :

$$\dot{K} = I - \delta K \quad (44.15)$$

The cost of investment is assumed to depend on the volume of investment as well as the rate of capital growth. The latter is due to adjustment cost: The faster capital grows, the more expensive is a given investment. A standard specification is a quadratic function for investment cost C :

$$C = I(1 + bI/K) \quad (44.16)$$

Parameter b measures the importance of adjustment costs. If investment per unit of capital goes to zero, goods are transformed into investment one to one, without adjustment cost.

Firms choose the level of investment at any moment such that the marginal investment cost equals the market stock price of capital q (so-called Tobin's q):

$$q = 1 + 2bI/K$$

Solving for I and inserting into Eq. (44.15) yields

$$\dot{K} = \begin{cases} K \left(\frac{q-1}{2b} - \delta \right) & \text{if } q \geq 1 \\ -\delta K & \text{else} \end{cases} \quad (44.17)$$

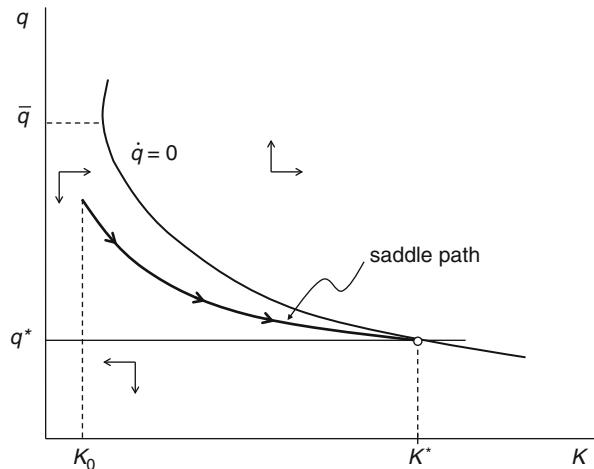
The stock price of capital q is the present value of future capital returns, similar as B in the dynamic migration model. Its dynamics follow from the non-arbitrage condition

$$\bar{r}q = \dot{q} - \delta q + f(K) + \frac{(q-1)^2}{4b} \quad (44.18)$$

In capital market equilibrium the interest on the market value of one unit of capital must equal the revaluation \dot{q} minus depreciation δq plus marginal productivity $f(K)$

²In a growth model with technical progress, f is increasing in time. For the sake of simplicity, we assume it to be time independent such that the long-term equilibrium is stationary.

Fig. 44.6 Tobin's q-theory of investment: In the course of time, the capital K and its stock price q in the region move along the saddle path until q is down at the stationary stock price q^* , and capital inflow ceases. As indicated by the arrows pointing into the direction of movement, the saddle path is the only path leading to the long-run equilibrium



plus the marginal contribution of capital to investment cost reduction according to Eq. (44.16).

With Eq. (44.18), solved for \dot{q} , and Eq. (44.15), we have two differential equations in the two variables K and q . Figure 44.6 depicts the dynamics in a phase diagram. The $(\dot{q} = 0)$ line is downward sloping for $q < \bar{q} = 1 + 2b(\bar{r} + \delta)$. The stationary state is (K^*, q^*) with

$$q^* = 1 + 2b\delta$$

and

$$K^* = f^{-1} \left((\bar{r} + \delta)q^* - \frac{(q^* - 1)^2}{4b} \right)$$

Note that, without adjustment costs ($b = 0$), we are back at the equilibrium condition with perfect real capital mobility:

$$\bar{r} = f(K^*) - \delta$$

If K_0 is smaller than K^* as in the figure, then $q(t_0) > q^*$. This induces positive net investment; capital grows and q steadily declines. If, due to some exogenous shock, the marginal productivity of capital goes up, the instantaneous response is not a capital “jump in,” but a “jump up” of the stock price q . This in turn lets investment instantaneously jump up and the capital starts growing steadily.

If $K_0 = K^*$ and, due to an exogenous shock, marginal productivity falls, then q falls below q^* . Hence, gross investment falls short of depreciation and the capital stock shrinks. If the downward shock is drastic enough, q even falls below one; investment ceases and capital declines at the depreciation rate δ .

44.7 Migration and Agglomeration

As mentioned, the assumption leading to the stabilizing role of factor migration is the neoclassical idea of decreasing marginal productivity. The so-called New Economic Geography (NEG) forcefully made the argument that this assumption cannot generally be true in a spatial economy (Krugman 1991). Otherwise it is impossible to explain the self-organization of the spatial economy such that agglomerations endogenously emerge on the one hand and sparsely populated areas on the other. If economies of scale in modern production sectors are strong enough and transportation costs small enough, then factor mobility does not lead to a uniform factor distribution in isotropic space, but on the contrary: If, in a thought experiment, obstacles to factor mobility are removed, a circular cumulative causation process sets in ending with the concentration of mobile factors at small spots that we call cities.

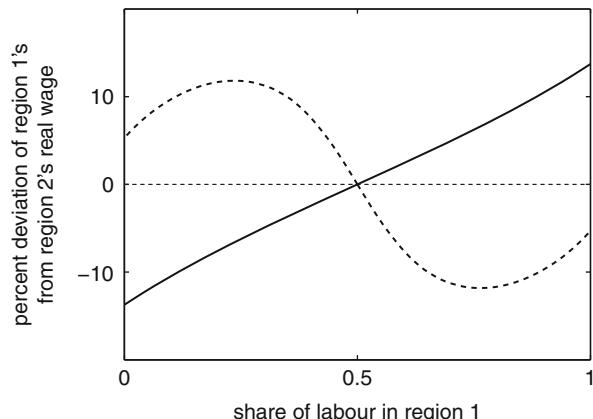
The emerging spatial structure is an equilibrium of centripetal and centrifugal forces. The former tend to make places with higher spatial concentration of mobile factors even more attractive for these factors. Two linkage effects generate the centripetal force: The *backward linkage effect* is due to the fact that factor movements to one region increase the size of the market in that region because the owners of the factors – provided they move with the factors – buy consumer goods. Firms and thus factors tend to follow the market because they want to save transport costs. Input and investment demand of firms also contribute to the backward linkage effect. The *forward linkage effect* is due to the fact that more factors in a region and thus higher supply available with low transport cost imply a lower price level in that region making it a more attractive location for even more factors to follow.

The centrifugal force is the *competition* or *crowding effect*. It is due to the fact that at least one factor is immobile. In Krugman's original center-periphery model, farmers are the immobile factor. As farmers are also consumers, the farmers' share in consumer demand has to be delivered to the farmers' location that is fixed and assumed to be distributed evenly across space. Firms thus also have an incentive not to be too far from the farmers' consumption demand.

To understand the role of factor mobility in NEG models we look at the so-called wiggle diagram in Fig. 44.7. It plots, for a NEG model with two symmetrical regions, the real wage difference (percentage deviation between region 1 and 2) over the distribution of the only mobile factor (labor in the nonfarm sector, called the “modern” sector) across the two regions. The basic assumptions of the model that the curve is derived from are the following:

- There are two regions (1 and 2) and 2 sectors, agriculture and “modern.”
- Agricultural goods are produced by immobile farmers under constant returns to scale.
- Modern goods are produced by mobile workers under increasing returns to scale.
- In the eyes of consumers, modern goods varieties are imperfect substitutes.
- Modern goods trade between regions is burdened with transport cost; agricultural goods trade is not. (The latter assumption is not essential, but made for the sake of tractability.)

Fig. 44.7 Wiggle diagram for the center-periphery model: Workers are attracted by higher real wages. The *solid* and *dotted* lines refer to low and high transport costs, respectively. The half-and-half distribution of labor across regions is stable (unstable) for high (low) transportation cost



The figure depicts two curves, the hump-shaped dotted curve for high, the solid monotone curve for low transport costs in the modern sector. For low transport costs the centripetal forces can be shown to dominate. The larger the share of the mobile factor in one region, the higher the wage in comparison to the other region. For high transport costs it is the other way round. The role of factor mobility is thus entirely different, depending on whether the economy is under the low transport cost or high transport cost regime. Assume labor is initially evenly distributed and now allowed to move. Under the high transport cost regime, nothing happens, while under the low transport cost regime, a slight random variation of the distribution makes the region with more workers more attractive. Hence, workers start moving to that region making it even relatively more attractive, until all workers reside in only one of the two regions, the endogenously emerging center. Which one of the two initially identical regions is going to become the center is a matter of historical coincidence.

While in Krugman's original model there is only one factor redistribution, from dispersed to concentrated, if transport costs decline from high to low, there are extended model versions with a second shift, back from concentrated to dispersed, if transport costs become low enough. The reason is a further immobile factor that is either an input of the modern sector or a consumption good. The higher the concentration of the mobile factor, the more its real returns get depressed by the scarcity of the immobile factor. When transport cost are low enough, the excess of the linkage effects over the competition effect, though still positive, is getting too small to compensate the negative effect of the scarce immobile factor (Puga 1999).

Distributional implications are similar to what we have seen in the neoclassical framework: Migrants gain, owners of immobile factors left behind in the periphery (farmers in this case) lose because they must pay more transport costs for consumer goods, and owners of immobile factors in the center gain because they have the modern sector closer by. In the extended model with an additional immobile input

factor in the modern sector, this input factor also gains in the center and loses in the periphery.

Questions about overall welfare gains or losses are much harder to answer. Because of imperfect competition no solution is Pareto optimal, and different allocations are usually Pareto incomparable because there are always winners and losers. One needs compensation criteria or a social welfare function for a comparison. No wonder, results in the literature are therefore rather diverse (see Behrens and Robert-Nicoud 2011, for a survey). In the basic model one finds that for low transport costs agglomeration is both a welfare optimum and market equilibrium. Therefore, if mobile factors are initially equally distributed among two symmetrical regions and mobility barriers are lifted, then they move to one of the regions, and this is desirable from a welfare point of view. If transport costs are high, then the dispersed equilibrium is both optimum and equilibrium. Initially equally distributed factors would not move, even if they could, and this is a desirable outcome. If, however, transport costs pass the so-called breakpoint from above, the dispersed equilibrium is replaced by a concentrated equilibrium. Initially equally distributed mobile factors start moving to one of the regions, though a dispersed equilibrium is still preferable if transport costs are lower than but sufficiently close to the breakpoint. Unfortunately, this conclusion is not robust against extensions of the model. In a model version where the modern sector needs an immobile factor, the market may also generate too much dispersion. Introducing additional centripetal forces can also lead to a market equilibrium with too little rather than too much agglomeration.

44.8 Conclusions

The previous sections show that factors tend to move to places where factor owners expect the highest returns. Under ideal circumstances factor movements enhance overall efficiency but also lead to considerable income redistribution such that not everyone is better off with free mobility rather than under a regime of restricted factor mobility. Under conditions of decreasing marginal productivity, factor movements support a dispersed factor distribution across space, but with increasing returns to scale and sufficiently low transport costs, they lead to concentration.

Though these are fairly clear conclusions, they are based on rather simplified concepts of migration decisions and migration incentives. Important branches of migration theory could not be dealt with in the previous sections due to space limitation, but shall briefly be mentioned. Some authors focus on the fact that migration decisions are not individual but *family decisions*. Many families look for a residential location offering jobs for more than one family member in an acceptable commuting distance. In this case the decision depends on the expected family income, taking migration costs as well as commuting costs in the destination into account. Beyond expected income, costs of educating children are an important migration motive (Mincer 1978). Family members may have diverging interests.

Thus, game-theoretic approaches help to explain the outcome of a family decision regarding migration.

An important issue in migration theory is *uncertainty*. Risk-averse individuals down-weight expected destination income if it is uncertain. Uncertainty is typically larger with regard to the destination region than the home region, such that a worker may prefer staying home in spite a net income gain to be expected in another region. Uncertainty also explains why migrants postpone a migration decision even though a move seems favorable. The reason is the *option value of waiting*. Potential migrants are facing a trade-off: An early move allows for reaping the benefit for a longer time but possibly lets the migrant miss the chance of staying or choosing a different destination after more information has become available. Finally, uncertainty is also among the factors explaining *destination clustering*. Cross-border migrants from one region of origin often cluster together in one region of destination. Closeness of friends and relatives or people with same language or culture eases information access. Sharing services like shops or cultural facilities also contributes to clustering.

Remittances are another important issue in international migration research. Remittances amount to more than 400 billion USD per annum and for some countries can be one third of GDP or more (World Bank 2012). The impact of remittances on the receiving countries is an active field of research beyond the scope of this chapter (Maimbo and Ratha 2005). Finally, migration and the welfare state have also only been touched upon above in the context of rural–urban migration. It is a wide field, also beyond the scope of this chapter.

References

- Acemoglu D (2009) Modern economic growth. Princeton University Press, Princeton
- Alonso W (1978) A theory of movements. Ballinger, Cambridge, MA, pp 197–211 (Chap 9)
- Barro R, Sala-i-Martin X (1995) Economic growth. McGraw-Hill, New York
- Behrens K, Robert-Nicoud F (2011) Tempora mutantur. J Econ Geogr 11(2):215–230
- Borjas GJ (1994) The economics of immigration. J Econ Lit 32(4):1667–1717
- Borjas GJ (2003) The labor demand curve is downward sloping: reexamining the impact of immigration on the labor market. Q J Econ 118:1335–1374
- Borjas GJ (2008) International migration. In: Durlauf SN, Blume LE (eds) The new Palgrave dictionary of economics. Palgrave Macmillan, Basingstoke
- Davies PS, Greenwood MJ, Li H (2001) A conditional logit approach to US state-to-state migration. J Reg Sci 41(2):337–360
- Glaeser EL (1999) Learning in cities. J Urban Econ 46(2):254–277
- Greenwood MJ (2007) Internal migration in developed countries. In: Rosenzweig MR, Stark O (eds) Handbook of population and family economics, vol 1B. Elsevier, Amsterdam, pp 647–720
- Harris JR, Todaro MP (1970) Migration, unemployment and development. Am Econ Rev 60(1):126–142
- Krugman P (1991) Increasing returns and economic geography. J Polit Econ 99(3):483–499
- Maimbo SM, Ratha D (2005) Remittances: development impact and future prospects. World Bank, Washington, DC
- Mincer J (1978) Family migration decisions. J Polit Econ 86(5):749–773

- Puga D (1999) The rise and fall of regional inequalities. *Eur Econ Rev* 43(2):303–334
- Sjaastad LA (1962) The costs and returns of human migration. *J Polit Econ Suppl* 70:80–89
- UNHCR (2010) Statistical yearbook. UNHCR, Geneva
- World Bank (2012) Payment systems and remittances: remittance market outlook. <http://web.worldbank.org>

Jan Oosterhaven and Geoffrey J. D. Hewings

Contents

45.1	Introduction	876
45.2	Interindustry Relations: The Base IO Table and the Demand-Driven IO Quantity Model	876
45.3	Adding Prices Without Interaction: The Dual Cost-Push Price Model	881
45.4	Adding Trade: The Interregional IO Table and Model	882
45.5	Adding Endogenous Consumption to the Interregional Model	891
45.6	Further Demo-economic Extensions of the Interregional IO Model	894
45.7	Conclusion	896
	Appendix: The Microeconomic Foundation of the Leontief and the Ghosh IO model	897
	References	900

Abstract

This chapter presents and critically evaluates the economic assumptions and applicability of a series of regional and interregional interindustry models. It begins with the demand-driven, single-region Leontief quantity model and its cost-push price dual. Then [Section 45.4](#) discusses the ideal, full information, interregional input–output model with interregional spillover and feedback effects at length, and compares it with the requirements and assumptions of more limited information, multiregional input–output models. [Section 45.5](#) discusses how to construct and add an interregional consumption function to obtain the type II interregional interindustry model. [Section 45.6](#) outlines further

J. Oosterhaven (✉)

Faculty of Economics and Business, University of Groningen, Groningen, The Netherlands
e-mail: j.oosterhaven@rug.nl

G.J.D. Hewings

Regional Economics Applications Laboratory, University of Illinois, Urbana-Champaign, IL, USA
e-mail: hewings@ad.uiuc.edu

extensions, all through to the most complex price-quantity interacting interregional demo-economic model LINE. Finally, an Appendix presents the microeconomic foundation for the Leontief model and compares it with the alternative supply-driven quantity model and its demand-pull price dual.

45.1 Introduction

The power of input–output (IO) analysis lies in its linking the sales of one industry, say agriculture, to the purchases of another industry, say the food industry. Interregional IO analysis adds a spatial dimension to this, for example, by linking the sales of the French agriculture to the purchases of the German food industry. In addition, input–output analysis places the data that describe these linkages into a single input–output table (IOT), such that it becomes directly clear that interregional IO analysis is based on a sectoral and regional disaggregation of the well-known macroeconomic equation for the gross regional or gross national product (GDP). With these data, a whole series of regional and interregional interindustry models may be built. This chapter discusses the assumptions of the basic version of these models, how these models can be solved, and what type of applications can be conducted with them.

We start in Sects. 45.2 and 45.3 with the basic IO quantity and the basic IO price model, developed by Wassily W. Leontief, who received the 1972 Nobel Prize for economics, especially for this work. In Sect. 45.4, we consider in more detail the ideal interregional IO model, developed by Walter Isard, the founding father of the Regional Science Association. In Sect. 45.5, the basic interregional IO model will be extended with an interregional consumption function. Section 45.6 will indicate how this model can be further extended and disaggregated, and how it can thus be used as a simple general equilibrium model, with prices and quantities interacting. Students with a strong theoretical interest will find the microeconomic foundation of the Leontief model, and a comparison with the alternative supply-driven IO model of Amica Ghosh (1958), in an Appendix.

45.2 Interindustry Relations: The Base IO Table and the Demand-Driven IO Quantity Model

Throughout this chapter, matrices are denoted by bold capitals, vectors by bold small types, and scalars by italics; \mathbf{x}' indicates the transpose of \mathbf{x} , $\hat{\mathbf{x}}$ a diagonal matrix of \mathbf{x} , \mathbf{i}' a summation row with ones, and $\mathbf{I} = \mathbf{1}\mathbf{1}'$ the identity matrix.

Figure 45.1 shows how the usual data in a national or *regional input–output table* are organized in four quadrants. The *first quadrant* contains the most salient data of the table, namely, the deliveries of intermediate products from industry i to industry j (indicated by z_{ij} , with $i, j = 1, \dots, N$). The *second quadrant* contains the deliveries of industry i to final demand category q (i.e., consumption, investments, government, and exports, indicated by y_{iq} , with $q = 1, \dots, Q$). The *third quadrant*

	Industry 1	Industry j	Industry N	Final demand	Total
Industry 1	.	1^{st} quadrant	.	2^{nd} quadrant	x_1
Industry i	.	z_{ij}	.	y_{iq}	x_i
Industry N	x_N
Imports		3^{rd} quadrant		4^{th} quadrant	M
Value added	.	v_{pj}	.	y_{pq}	Y
Total	x_1	x_j	x_N	$C \quad I \quad G \quad E$	

Fig. 45.1 Regional input–output table, with four quadrants and macroeconomic totals. Legend: z_{ij} , intermediate deliveries from industry i to industry j ; y_{iq} , final deliveries from industry i to final demand type q ; x_i , total output/input of industry i ; v_{pj} , primary input type p purchased by industry j ; y_{pq} , primary inputs type p purchased by final demand category q ; C , household consumption; I , investments; G , government expenditures; E , exports; M , imports; Y , gross value added at market prices

contains the primary inputs of category p (i.e., imports and the various components of gross value added at market prices) used by industry j (indicated by v_{pj} , with $p = 1, \dots, P$). The *fourth quadrant* contains the primary inputs of type p that are purchased by final demand category q (y_{pq}). The most important of these are the imports of consumption and investment goods.

The row totals of the first and second quadrant (x_i) equal total sales by industry i , which is made equal to total output by including *changes in stocks* as part of investments. Calculating percentages across these rows enables interesting analyses of the differences in market and *sales structure* of the industries distinguished. The column totals of the first and third quadrant (x_j) equal total cost of industry j , which is made equal to total output by including the *net operating surplus* as part of the gross value added in market prices. Calculating percentages across these columns allows for comparative analyses of the purchase and *cost structure* of various industries. Since these row and column totals are equal by industry, the overall total of the third and fourth quadrant ($M + Y$) and the overall total of the second and fourth quadrant ($C + I + G + E$) are also equal. The rearrangement of these totals shows that an IOT, in fact, represents a sectorally detailed view of the well-known *macroeconomic identity* for the gross regional or gross national product/income (GDP), namely, $Y = C + I + G + E - M$.

Besides descriptive statistical analyses of sales and cost structures, an input–output table also provides the data to specify a series of interindustry models. The accounting identities of these models are usually based on those of an IOT. Additionally, these models require behavioral and institutional assumptions, and assumptions about which variables are determined outside the model (called exogenous variables), and which are determined inside the model (called endogenous variables). We start our exposition of these models with the most simple and oldest of them (Leontief 1936).

This *demand-driven IO quantity model* is based on the accounting identities for the rows of an IOT for a *closed economy*, that is, for Fig. 45.1 without the import row and without the export column. This model has two core behavioral assumptions. The first stipulates that the supply of output of all industries, $i = 1, \dots, N$, follows the total of the intermediate demands z_{ij} and the total of the final demands f_{iq} for its products:

$$x_i = \sum_j z_{ij} + \sum_q y_{iq} \text{ for all } i \quad (45.1)$$

or in matrix algebra : $\mathbf{x} = \mathbf{Z} \mathbf{i} + \mathbf{Y} \mathbf{i} = \mathbf{Z} \mathbf{i} + \mathbf{y}$

where the N -by- N matrix \mathbf{Z} represents the first quadrant, the N -by- Q matrix \mathbf{Y} the second quadrant, and the N -column \mathbf{x} the row totals of the first plus second quadrant of Fig. 45.1. Note that these three types of quantities are all defined as unit quantities, with a constant price equal to one (not shown explicitly), such that they may be summed by row and column. Thus, Eq. (45.1) assumes that supply follows demand without any price change or stimulus. This means that each industry's supply is infinitely price elastic, which is a plausible assumption in short run situations with spare production capacity, or in long run situations in which the relative prices of the inputs on the supply side do not change.

The second behavioral assumption is that the demand for intermediate inputs z_{ij} and primary inputs v_{pj} is linearly and solely determined by the total output of purchasing industry j :

$$z_{ij} = a_{ij}x_j, \quad \text{for all } i, j \quad \text{or in matrix algebra : } \mathbf{Z} \mathbf{i} = \mathbf{A} \mathbf{x} \quad (45.2a)$$

$$v_{pj} = c_{pj}x_j, \quad \text{for all } p, j \quad \text{or in matrix algebra : } \mathbf{V} \mathbf{i} = \mathbf{C} \mathbf{x} \quad (45.2b)$$

where the P -by- N matrix \mathbf{V} represents the third quadrant of Fig. 45.1. Note that the assumption of constant prices equal to one, which is implicitly present in Eq. (45.2a, b), implies that the demand for intermediate and primary inputs has a price inelasticity of zero, whereas the supply of intermediate and primary inputs is perfectly price elastic. Taken together, these assumptions imply that there are no bottlenecks in the region's labor, land, or capital markets.

The *technical coefficients* a_{ij} and c_{pj} in Eq. (45.2a, b) indicate, respectively, the amount of intermediate inputs from industry i , and the amount of primary inputs of category p , needed per unit of output of industry j . When only one IOT is available, the matrices \mathbf{A} and \mathbf{C} are simply estimated by the column-wise division of each element of the intermediate inputs matrix \mathbf{Z} , and the primary inputs matrix \mathbf{V} , by the total of the corresponding column of the IOT. In such cases, $\mathbf{A} = \mathbf{Z} \hat{\mathbf{x}}^{-1}$ and $\mathbf{C} = \mathbf{V} \hat{\mathbf{x}}^{-1}$, with the overall column total of the technical coefficients being equal to one, that is, $\mathbf{i}' \mathbf{A} + \mathbf{i}' \mathbf{C} = \mathbf{i}'$. Note that this specification implies that there are *no* economies of scale, while all intermediate and primary inputs are mutually complementary.

Fig. 45.2 The causal structure of the basic IO quantity model

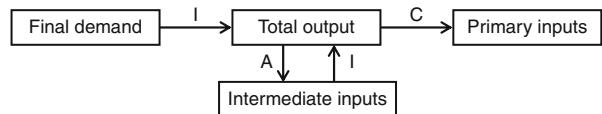


Figure 45.2 summarizes the causal structure of the basic IO quantity model. The symbols next to the arrows indicate the size of the *direct effect* along the direction of the arrow. The symbol **A**, for instance, indicates that a change in the total output vector (Δx) leads to a direct change in intermediate input matrix **Z** that is equal to **A** times Δx , while the arrows with the symbol **I** indicate a one-to-one backward impact of demand on the corresponding supply.

Figure 45.2 shows that the demand for final outputs y is exogenous, as no arrows are coming in. Any change in y will lead to an equally large direct change of $\mathbf{I} \Delta y$ in total output x . This change in total output, in its turn, will lead to *first round indirect effects* on the demand for intermediate and primary inputs of, respectively, $\mathbf{A} \Delta y$ and $\mathbf{C} \Delta y$. The 1st round effect on primary inputs will lead to *no* further changes in any of the endogenous variables, as no arrows go out. The 1st round effect on intermediate inputs, however, will lead to an equally large backward change in total output, indicated by **I**, which will lead to *second round indirect effects* on the demand for intermediate and primary inputs of, respectively, $\mathbf{A}^2 \Delta y$ and $\mathbf{C} \mathbf{A} \Delta y$. The third round indirect effects amount to, respectively, $\mathbf{A}^3 \Delta y$ and $\mathbf{C} \mathbf{A}^2 \Delta y$, and so on. Consequently, the equilibrium size of total output equals

$$\mathbf{x} = (\mathbf{I} + \mathbf{A} + \mathbf{A}^2 + \mathbf{A}^3 + \dots) \mathbf{y} = (\mathbf{I} - \mathbf{A})^{-1} \mathbf{y} = \mathbf{L} \mathbf{y} \quad (45.3)$$

If all column sums of **A** are smaller than one, which implies that value added is positive for each industry, the *Taylor expansion* of the **A** matrix converges to the so-called *Leontief inverse* $\mathbf{L} = (\mathbf{I} - \mathbf{A})^{-1}$ in Eq. (45.3). However, since the input-output model is a comparative static equilibrium model, it does not specify either the length or the nature of the equilibrium process. Most IO applications, in fact, work with year-to-year changes of one equilibrium to the next. When large shocks to the economy need to be modeled, it may be necessary to assume a longer period before the new equilibrium is reached.

The equilibrium solution for endogenous total output in Eq. (45.3) may also be found by substituting Eq. (45.2a) in Eq. (45.1), transferring **A** \mathbf{x} to the left-hand side, and pre-multiplying both sides by the Leontief inverse. The equilibrium solutions for endogenous intermediate and primary inputs, in their turn, are found by substituting Eq. (45.3) in Eqs. (45.2a) and (45.2b), respectively, yielding

$$\mathbf{Z} \mathbf{i} = \mathbf{A}(\mathbf{I} - \mathbf{A})^{-1} \mathbf{y} = \mathbf{A} \mathbf{L} \mathbf{y} \quad (45.4a)$$

$$\mathbf{v} = \mathbf{V} \mathbf{i} = \mathbf{C}(\mathbf{I} - \mathbf{A})^{-1} \mathbf{y} = \mathbf{C} \mathbf{L} \mathbf{y} \quad (45.4b)$$

with $\mathbf{v} = P$ -vector with the economy-wide total primary inputs of type p , that is, the row totals of the third quadrant of Fig. 45.1. Equations (45.3) and (45.4a, b) specify how the endogenous variables \mathbf{x} , \mathbf{Z} and \mathbf{v} depend on exogenous final demand \mathbf{f} through output, and intermediate and primary input multipliers, respectively.

The IO literature pays much attention to the *output multipliers* from the Leontief inverse $(\mathbf{I} - \mathbf{A})^{-1}$. Its typical element, l_{ij} , indicates the direct and indirect need for outputs of industry i per unit of final demand for products of industry j , while the column sums of the Leontief inverse indicate the economy-wide total output effect of the same unit of final demand. For policy purposes, however, the *employment multipliers* and the *income multipliers* from matrix $\mathbf{C} \mathbf{L}$ in Eq. (45.4b) are much more interesting. If p relates to the use of labor or total value added, its typical element, $cl_{pj} = \sum_i c_{pi} l_{ij}$, indicates the economy-wide direct and indirect employment or income per unit of final demand for products from industry j .

Such primary input multipliers are used in a whole array of applications. The most common of these are all kind of estimations of the income, employment, or CO₂ emissions embodied in, for example, consumption, investments, or exports. In the case of CO₂ emissions or energy use, the matrix \mathbf{V} is simply replaced with a single row of emission levels or energy use by industry, \mathbf{v}' , and the matrix \mathbf{C} is simply replaced with a row with CO₂ *emission coefficients* or *energy use coefficients* per unit of industry output, \mathbf{c}' . Input–output practitioners should be wary about the implications of the assumptions underlying, especially, the income and employment multipliers, as authorities and firms will press for large multipliers to serve their lobby needs. If, for instance, the regional labor market is tight, the impact of a demand shock may materialize in an increase of local wages, instead of the job growth that is predicted by the IO employment multiplier.

Policy makers have been fascinated by multipliers, but they often neglect differences in quality (e.g., skill levels) of jobs and focus instead on the size of the multipliers. Further, there is often confusion about the interpretation of an employment multiplier since the indirect effects may involve parts of many hundreds or thousands of jobs that when netted out generate only a modest number. This will typically be the case for *impact analysis* of short-term events; for example, participants in the Chicago Marathon usually stay 2–3 nights in the region. While there, they will spend money in restaurants generating part of the daily income to many hundreds of waiters – but only for the period in which they are in the region. The resulting multipliers might reveal 50 full-time equivalent jobs in the restaurant sector from the impacts of their spending, but in reality, hundreds of parts of jobs will have been affected.

Further, as transportation costs have decreased in real terms, interregional trade has increased sharply; this has resulted in a *hollowing out* of many regional economies as intra-regional purchases are replaced by interregional ones. The resulting intra-regional multipliers have often decreased, generating concern among policy makers that the region may be losing its competitiveness. Spurred on by the promise of cluster-based development strategies, underpinned by IO analyses of existing *clusters of industries*, there is a concomitant expectation that multipliers should increase over time. However, the competitiveness of each component of a supply

chain may see production systems having a much more extensive geographical imprint; thus, both the magnitude and composition (intra- versus interregional or feedback effects) of multipliers are likely to change over time.

45.3 Adding Prices Without Interaction: The Dual Cost-Push Price Model

As said above, prices do not play a role in the IO quantity model, but they assume center stage in the dual *cost-push IO price model* (Leontief 1951). The causal structure of that model is also shown in Fig. 45.2, but then the arrows should be imaged as running in the opposite, forward direction, while the boxes then refer to prices, and not to quantities.

This mirror image of Fig. 45.2 shows that the P prices of the primary inputs (p_p), along the rows of the third quadrant of Fig. 45.1, are the exogenous variables in the IO price model, as none of the reversed arrows is coming in. Any change in the P -column with these prices (\mathbf{p}_v) leads to a change in N -column with the output prices per industry (\mathbf{p}'), the size of which is determined by the P -by- N matrix with the *cost shares* of these primary inputs in total output (\mathbf{C}). Each change in a single output price, in its turn, is entirely passed on to all intermediate and final users of that output, as indicated by the \mathbf{I} matrix. In the case of the final users, this leads to no further changes, as no reversed arrows are going out. In the case of intermediate users (i.e., firms), however, any change in their intermediate input prices is passed on to the firms that use their outputs, as indicated by the \mathbf{I} matrix. This leads to a further forward change in output prices, the size of which is determined by matrix with the cost shares of the intermediate inputs in total output (\mathbf{A}). Hence, the IO price model is very suited to model forward, cost-push effects of primary input prices on final output prices.

The mathematics of the IO price model formalizes the above explanation. Its accounting identities are based on the columns of the IOT, instead of on its rows, as in the quantity model. Moreover, now prices are made explicit, whereas in the quantity model, they were implicit, held constant, and set equal to one. The accounting identities for the *values* of the columns of the IOT equal:

$$p_j x_j = \sum_i p_i z_{ij} + \sum_p p_p v_{pj}, \text{ for all } j \quad (45.5)$$

or in matrix algebra : $\mathbf{p}' \hat{\mathbf{x}} = \mathbf{p}' \mathbf{Z} + \mathbf{p}_v' \mathbf{V}$

Substitution of Eqs. (45.2a) and (45.2b) in Eq. (45.5), and post-multiplication with $\hat{\mathbf{x}}^{-1}$, reveals the accounting identities for the total output prices, which equal the sum of their intermediate and primary input prices weighted by their corresponding cost shares:

$$\mathbf{p}' = \mathbf{p}' \mathbf{A} + \mathbf{p}_v' \mathbf{C} \quad (45.6)$$

Adding the assumption that all price changes are entirely and precisely passed on to all users makes it possible to solve for the final output prices \mathbf{p}' as a function of the primary input prices \mathbf{p}_v' :

$$\mathbf{p}' = \mathbf{p}_v' \mathbf{C}(\mathbf{I} - \mathbf{A})^{-1} \quad (45.7)$$

Note that the *output price multipliers* of the primary input prices, $\mathbf{C}(\mathbf{I} - \mathbf{A})^{-1}$ in Eq. (45.7), are equal to the primary input multipliers of final demand in Eq. (45.4b). Also note that the column sum of these price multipliers is equal to one, as $\mathbf{i}'\mathbf{C} + \mathbf{i}'\mathbf{A} = \mathbf{i}'$ implying that $\mathbf{i}'\mathbf{C} = \mathbf{i}'(\mathbf{I} - \mathbf{A})$. Both observations make sense as both types of multiplier show the amount of primary inputs (capital, labor and land) embodied in final output.

This primal-dual model relationship becomes even more evident when Eq. (45.7) is post-multiplied with final demand \mathbf{y} , which provides the following expression:

$$\mathbf{p}'\mathbf{y} = \mathbf{p}_v' \mathbf{C}(\mathbf{I} - \mathbf{A})^{-1}\mathbf{y} = \mathbf{p}_v' \mathbf{v} \quad (45.8)$$

Equation (45.8) shows that the value of total final demand equals the value of total primary inputs, as already shown by $C + I + G + E = M + Y$ in Fig. 45.1. In the case of the IO quantity model, the focus of the model behind Eq. (45.8) is on the *backward causal impact* of any change in final demand on the, directly and indirectly, necessary primary inputs. In the case of the IO price model, the emphasis is on the *forward causal impact* of any change in primary input prices, both directly and indirectly, on the prices paid by the final users (consumers, investors, and government).

Typical applications of the IO price model inform policy makers on such questions as the impact of oil price hikes on consumer prices, or on the impact CO₂ taxes or production subsidies on the competitive price of exports. In a very early interregional application, Oosterhaven (1981a) used an interregional IO price model, with an endogenous instead of exogenous wage rate, to simulate the regionally different consumer price impacts of the increases in the oil and natural gas prices in the 1970s, in the case of the Netherlands. To be able to repeat such an analysis, we first have to add space and endogenous consumption to the standard IO price and quantity model, which is the topic of the next two sections.

45.4 Adding Trade: The Interregional IO Table and Model

Figure 45.3 shows the setup of the so-called “ideal” *interregional input-output table* (IRIOT), devised by Isard (1951), with the IO data for a national economy split-up into R regions and N industries. The first quadrant contains the NR -by- NR intermediate demand block matrix \mathbf{Z} , with \mathbf{Z}^{rs} as its typical square block showing the sales of the N industries in region r to their N sister industries in region s .

	Industry demand			Final demand				Total
	Region 1	...	Region R	Region 1	...	Region R	Foreign exports	
Region 1	Z^{11}	...	Z^{1R}	F^{11}	...	F^{1R}	e^1	x^1
:	:		Z^{rs}	:		F^{rs}	e^r	x^r
Region R	Z^{R1}	...	Z^{RR}	F^{R1}	...	F^{RR}	e^R	x^R
Foreign imports	Z^{m1}	...	Z^{mR}	F^{m1}	...	F^{mR}	Transit trade	M^{for}
Value added	V^1	...	V^R	Y^1	...	Y^R	$\mathbf{0}$	Y^{nat}
Total	x^1	x^s	x^R	$C^1 I^1 G^1$...	$C^R I^R G^R$	E^{for}	

Fig. 45.3 The “ideal” interregional input–output table for R regions. Legend: see Fig. 45.1. All double-superscripted matrices relate to interregional trade from the origin region (*first superscript*) to the destination region (*second superscript*)

The distinction between the diagonal and the off-diagonal blocks of Z is crucial. The diagonal blocks show the *intra-regional deliveries* of intermediate goods and services, z_{ij}^{rr} , whereas the off-diagonal blocks show the *interregional trade* in intermediate goods, z_{ij}^{rs} , from industry i in r (note: first indices = origin) to industry j in s (note: second indices = destination).

The second quadrant contains the final demand block matrix that consists of the domestic final demand block matrix F and a foreign export block column e . The typical N -by- Q rectangular block F^{rs} contains the sales of final goods and services by the N industries in region r to the consumers, investors, and government in region s . Again, the distinction between the diagonal and the off-diagonal blocks is important. The diagonal blocks show the intra-regional deliveries of final goods and services to demand category q within region r , f_{iq}^{rr} , whereas the off-diagonal blocks show the interregional trade of final goods from industries in r to consumers, investors, and government in s , f_{iq}^{rs} . As for the block column e , note that its typical column e^r contains the foreign exports of both *intermediate* goods, for the industries, and *final* goods, for the consumers, investors, and government, in the Rest of the World (Row).

The third quadrant contains the primary input block matrix, which consist of the N -by- N square foreign import blocks Z^{ms} and the P -by- N rectangular value added blocks V^s . The foreign import blocks contain the imports of products from industry i in the RoW by the industries j in s (z_{ij}^{ms}). The value added blocks contain the usual components of gross value added at market prices (production taxes less subsidies, gross wages, employers’ contributions to social security, and the operating surplus) of industries j in s (v_{pj}^s). The fourth quadrant again contains the primary inputs of final demand.

As in the case of Fig. 45.1, the total output along the rows of Fig. 45.3 equals the total input along its columns. Consequently, total final demand

$(\sum_r C^r + \sum_r I^r + \sum_r G^r + E^{for})$ again equals total primary input ($Y^{nat} + M^{for}$). In Sect. 45.2, it was shown that a national IOT represents a sectoral disaggregation of the *macroeconomic identity*: $Y = C + I + G + E^{for} - M^{for}$. Here, it becomes clear that an interregional IOT represents an additional, regional disaggregation of the same identity. Moreover, the rearrangement of the elements of Fig. 45.3 shows that an IRIOT also includes all (but now much more detailed) identities for gross regional product/income:

$$\begin{aligned} Y^r &= \mathbf{i}' \mathbf{V}^r \mathbf{i} + \mathbf{i}' \mathbf{Y}^r \mathbf{i} = C^r + I^r + G^r \\ &+ \left(\sum_{s \neq r} \mathbf{i}' \mathbf{Z}^{rs} \mathbf{i} + \sum_{s \neq r} \mathbf{i}' \mathbf{F}^{rs} \mathbf{i} + \mathbf{i}' \mathbf{e}^r \right) \\ &- \left(\sum_{s \neq r} \mathbf{i}' \mathbf{Z}^{sr} \mathbf{i} + \sum_{s \neq r} \mathbf{i}' \mathbf{F}^{sr} \mathbf{i} \right) = C^r + I^r + G^r + E^r - M^r \end{aligned} \quad (45.9)$$

Note that the $\sum_{s \neq r}$ in Eq. (45.9) is needed to exclude the intra-regional transactions from total regional exports (the first term between brackets) as well as from total regional imports (the second term between brackets).

Calculating percentages across the rows of the IRIOT enables interesting comparative analyses of interregional sales structures of industries across different regions, whereas calculating percentages across the columns allows for an analysis of cost structures and interregional purchases structures of industries across regions. The main use of an IRIOT, however, is to supply the accounting identities of the various interregional IO models and the coefficients for the behavioral equations of such models.

The mathematics of the basic *interregional IO quantity model* is rather similar to that of the single-region IO model of Eqs. (45.1) and (45.2a, b). Written with all blocks separately, the interregional IO model for an open national economy reads as follows:

$$\mathbf{x} = \begin{bmatrix} \mathbf{x}^1 \\ \vdots \\ \mathbf{x}^R \end{bmatrix} = \begin{bmatrix} \mathbf{A}^{11} & \cdots & \mathbf{A}^{1R} \\ \vdots & \ddots & \vdots \\ \mathbf{A}^{R1} & \cdots & \mathbf{A}^{RR} \end{bmatrix} \begin{bmatrix} \mathbf{x}^1 \\ \vdots \\ \mathbf{x}^R \end{bmatrix} + \begin{bmatrix} \mathbf{F}^1 \mathbf{i} \\ \vdots \\ \mathbf{F}^R \mathbf{i} \end{bmatrix} + \begin{bmatrix} \mathbf{e}^1 \\ \vdots \\ \mathbf{e}^R \end{bmatrix} \quad (45.10a)$$

$$\mathbf{v} = \begin{bmatrix} \mathbf{v}^1 \\ \vdots \\ \mathbf{v}^R \end{bmatrix} = \begin{bmatrix} \mathbf{c}^1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \mathbf{c}^R \end{bmatrix} \begin{bmatrix} \mathbf{x}^1 \\ \vdots \\ \mathbf{x}^R \end{bmatrix} \quad (45.10b)$$

Using the block matrices of Fig. 45.3, its solution for total output and value added is similar to Eqs. (45.3) and (45.4):

$$\mathbf{x} = (\mathbf{I} - \mathbf{A})^{-1} (\mathbf{F} \mathbf{i} + \mathbf{e}) = \mathbf{L}^* (\mathbf{F} \mathbf{i} + \mathbf{e}) \quad (45.11a)$$

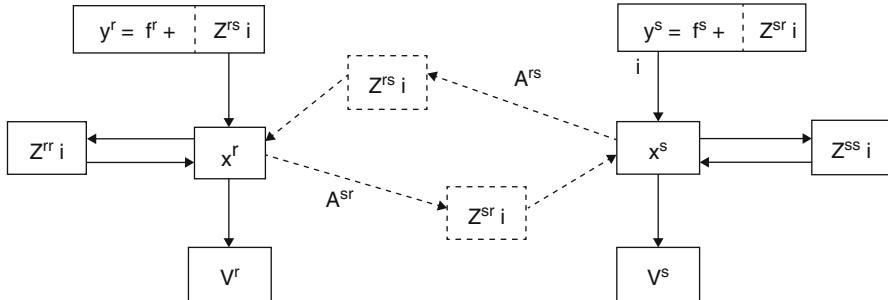


Fig. 45.4 The causal structure of the interregional IO model extension. Legend: y , vector with exogenous final demand of single-region IO model by sector of origin; f , vector with exogenous demand of interregional IO model by sector of origin; Z^{rs} , interindustry matrix with intermediate exports from region r to region s ; x , vector with total output by sector; V , matrix with value added type, by sector

$$\mathbf{v} = \hat{\mathbf{c}}(\mathbf{I} - \mathbf{A})^{-1}(\mathbf{F} \mathbf{i} + \mathbf{e}) = \hat{\mathbf{c}}\mathbf{L}^* (\mathbf{F} \mathbf{i} + \mathbf{e}) \quad (45.11b)$$

where the interregional input coefficient matrices \mathbf{A} and $\hat{\mathbf{c}}$ may be derived from an IRIOT, like Fig. 45.3, by means of $\mathbf{A} = \mathbf{Z} \hat{\mathbf{x}}^{-1}$ and $\mathbf{c}' = \mathbf{v}' \hat{\mathbf{x}}^{-1}$, and where \mathbf{L}^* denotes the *interregional Leontief inverse*, whose typical element, l_{ij}^{rs} , indicates the direct and indirect output from industry i in r needed per unit of final demand for the outputs of industry j in s . Analogously, the *interregional income multiplier* matrix $\hat{\mathbf{c}} \mathbf{L}^*$ has c_i^{rs} as its typical element, indicating the direct and indirect value added in industry i in r needed per unit of final demand for the product of industry j in s . Although the mathematical structure of the basic interregional IO model of Eq. (45.10a, b) is rather similar to that of the single-region model, its economic interpretation and behavioral implications are more complex.

To sharpen our insight, we first compare the causal structure of the interregional model with only two regions r and s with the single-region model, by means of Fig. 45.4. For reasons of simplicity, Fig. 45.4 assumes that the two regions r and s together form a *closed economy*, that is, there are no foreign imports or foreign exports. The bold lines and boxes show the causal structure of the two independent single-region models, that is, they represent a double version of Fig. 45.2. The dotted separation in the two top boxes indicates which part of the single-region's final demand remains exogenous, and which part is made endogenous by adding the two dotted interregional trade boxes that link the two independent single-region models into one interregional model.

From the viewpoint of region r , its formerly exogenous exports of intermediate goods to region s are now endogenously determined by the output levels of the purchasing industries in region s , that is, $\mathbf{Z}^{rs}\mathbf{i} = \mathbf{A}^{rs}\mathbf{x}^s$, while the formerly exogenous intermediate exports of s are now explained as imports of region r 's industries, that is, $\mathbf{Z}^{sr}\mathbf{i} = \mathbf{A}^{sr}\mathbf{x}^r$. Consequently, the exogenous demand for the outputs of both regions has shrunken. Their output levels, however, do not change.

The same reality is only modeled in a different way! Thus, the smaller size of exogenous demand has to be neutralized by larger multipliers. And, indeed, the multipliers of the interregional IO model are larger because they add the *interregional feedbacks effects*, from region r via region s back to region r , to the single-region model. Following the dotted arrows of Fig. 45.4 shows that the interregional feedback effects for region r , actually, consist of two *interregional spillover effects*, the first from r to s and the second back from s to r , enhanced by the *intra-regional multipliers* of region s . In more formal terms, the interregional feedback effects of region r 's final demand on its own output levels equal $\mathbf{A}^{rs}(\mathbf{I} - \mathbf{A}^{ss})^{-1}\mathbf{A}^{sr}$.

This interregional feedback formula may be derived mathematically, by writing out and solving the partitioned version of the *two-region IO model* of Fig. 45.4:

$$\mathbf{x}^r = \mathbf{Z}^{rr}\mathbf{i} + \mathbf{Z}^{rs}\mathbf{i} + \mathbf{f}^r = \mathbf{A}^{rr}\mathbf{x}^r + \mathbf{A}^{rs}\mathbf{x}^s + \mathbf{f}^r \quad (45.12a)$$

$$\mathbf{x}^s = \mathbf{Z}^{sr}\mathbf{i} + \mathbf{Z}^{ss}\mathbf{i} + \mathbf{f}^s = \mathbf{A}^{sr}\mathbf{x}^r + \mathbf{A}^{ss}\mathbf{x}^s + \mathbf{f}^s \quad (45.12b)$$

The first equalities of Eq. (45.12a, b) show the accounting identities across the rows of the underlying bi-regional IO table. The second equalities show how the intra- and interregional input coefficients together with the output levels of the purchasing industries determine the endogenous domestically produced inputs and the endogenous imported inputs. A step-by-step solution of Eq. (45.12a, b) gives the disaggregated solution of the two-region IO model:

$$\begin{aligned} \mathbf{x}^r &= \left[\mathbf{I} - \mathbf{A}^{rr} - \mathbf{A}^{rs}(\mathbf{I} - \mathbf{A}^{ss})^{-1}\mathbf{A}^{sr} \right]^{-1} \left[\mathbf{f}^r + \mathbf{A}^{rs}(\mathbf{I} - \mathbf{A}^{ss})^{-1}\mathbf{f}^s \right] \\ &= \mathbf{L}^{rr}\mathbf{f}^r + \mathbf{L}^{rs}\mathbf{f}^s \end{aligned} \quad (45.13a)$$

$$\begin{aligned} \mathbf{x}^s &= \left[\mathbf{I} - \mathbf{A}^{ss} - \mathbf{A}^{sr}(\mathbf{I} - \mathbf{A}^{rr})^{-1}\mathbf{A}^{rs} \right]^{-1} \left[\mathbf{f}^s + \mathbf{A}^{sr}(\mathbf{I} - \mathbf{A}^{rr})^{-1}\mathbf{f}^r \right] \\ &= \mathbf{L}^{sr}\mathbf{f}^r + \mathbf{L}^{ss}\mathbf{f}^s \end{aligned} \quad (45.13b)$$

Equations (45.13a) and (45.13b) may also be written with block matrices and vectors, which explicitly shows the structure of the interregional Leontief inverse \mathbf{L}^* for two regions:

$$\begin{aligned} \mathbf{x} &= \begin{bmatrix} \mathbf{x}^r \\ \mathbf{x}^s \end{bmatrix} \\ &= \begin{bmatrix} \left[\mathbf{I} - \mathbf{A}^{rr} - \mathbf{A}^{rs}(\mathbf{I} - \mathbf{A}^{ss})^{-1}\mathbf{A}^{sr} \right]^{-1} & \mathbf{L}^{rr}\mathbf{A}^{rs}(\mathbf{I} - \mathbf{A}^{ss})^{-1} \\ \mathbf{L}^{ss}\mathbf{A}^{sr}(\mathbf{I} - \mathbf{A}^{rr})^{-1} & \left[\mathbf{I} - \mathbf{A}^{ss} - \mathbf{A}^{sr}(\mathbf{I} - \mathbf{A}^{rr})^{-1}\mathbf{A}^{rs} \right]^{-1} \end{bmatrix} \begin{bmatrix} \mathbf{f}^r \\ \mathbf{f}^s \end{bmatrix} \\ &= \mathbf{L}^*\mathbf{f} \end{aligned} \quad (45.13)$$

Thus, total output of region r is determined by the exogenous part of its own final demand \mathbf{f}^r as well as by the exogenous part of the final demand of the other region \mathbf{f}^s .

There are two differences between the multipliers of a single-region IO model compared to those of an interregional IO model. The first difference shows up in the intra-regional impact of the own final demand \mathbf{f}^r on the own output \mathbf{x}^r . In the single-region model, it equals $(\mathbf{I} - \mathbf{A}^{rr})^{-1} \mathbf{f}^r$, whereas in the interregional model, the same effect equals $[\mathbf{I} - \mathbf{A}^{rr} - \mathbf{A}^{rs} (\mathbf{I} - \mathbf{A}^{ss})^{-1} \mathbf{A}^{sr}]^{-1} \mathbf{f}^r$. Clearly, the last impact is larger than the first, as it includes the *interregional feedbacks* of the own final demand via the own imports on the exports and output of the other region, and the subsequent effect of the imports of the other region back on the own region's exports. These larger intra-regional multipliers are, however, compensated by the smaller size of the exogenous demand in the interregional model ($\mathbf{f}^r = \mathbf{y}^r - \mathbf{Z}^{rs} \mathbf{i}$) compared to the single-region model (\mathbf{y}^r).

The empirical size of this first difference has been studied extensively (see Miller and Blair 2009). Unfortunately, it is almost always measured by dividing the interregional feedback effect by the total intra-regional effect, thus including the one-to-one direct effect of final demand on total output, as captured in the \mathbf{I} matrix of the Taylor expansion of the Leontief inverse in Eq. (45.3). However, neither the IO model nor any other model is needed to estimate this direct effect. Hence, the underestimation of the intra-regional impact should only be judged by the indirect part of the impact, as captured by the $(\mathbf{L} - \mathbf{I})$ matrix.

In the case of the Dutch economy in 1970, measured for the indirect impacts only, Oosterhaven (1981b) found an underestimation of the regional income effects of regional final demand of only 1.1 % for the relatively isolated rural Northern Netherlands and a small 3.4 % for the heavily urbanized greater Rotterdam region. When type II multipliers, with endogenous consumption expenditures, as discussed in the next section, were used, the neglect of interregional feedbacks led to a larger underestimation of the regional income impacts of 3.1 % and 6.6 %, respectively. The reason for the larger feedbacks was that interregional commuting and interregional shopping effects were included in the type II multipliers, whereas they are absent in the basic interregional IO model.

The second difference between the two models is hardly discussed in the literature but is at least as important. It shows up in the interregional spillover effect of the own final demand \mathbf{f}^r on the output of the other region \mathbf{x}^s . In the single-region IO model, the interregional spillover is measured by $\mathbf{A}^{sr} (\mathbf{I} - \mathbf{A}^{rr})^{-1}$, whereas it equals $\mathbf{L}^{ss} \mathbf{A}^{sr} (\mathbf{I} - \mathbf{A}^{rr})^{-1}$ in Eq. (45.13c). Hence, the *interregional spillovers* are also larger when estimated with the interregional IO model than when estimated with the single-region model. The difference being, of course, that the interregional model takes the intra-regional effects inside the other region into account, as shown by the added \mathbf{L}^{ss} , whereas the single-region model is unable to do so.

Recent research for the 27 members of the European Union (EU) analyzed the differences in the EU27 income effects of the EU27 exports to third countries, as estimated with 27 separate national IO models and as estimated with a single

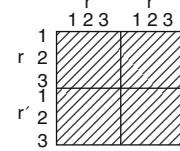
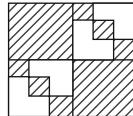
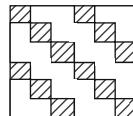
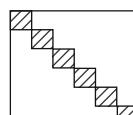
consolidated IO model for the EU27 as a whole (Bouwmeester, et al. 2012). It reports a weighted average *first round intra-EU income spillover* in the rest of the EU27, as calculated with the 27 single-country models, of 7.7 % of the domestic income effect of the country at hand. The additional *higher-order intra-EU income spillovers and feedback effects*, as calculated with the full EU27 model, appeared to be as large as 10.7 % of the weighted average domestic effect. The relatively large size of the higher-order effects, in this case, must be attributed to the fact that they relate to the interactions between as much as 27 countries. Hence, the underestimation of interregional spillover effects with a single-region IO model seems to be much more serious issue than the much discussed underestimation of the intra-regional effect.

One last theoretically important aspect of the basic interregional input–output (IRIO) model needs to be discussed, as it applies to all interregional IO models. In the closed economy model of Sect. 45.2, the input coefficients a_{ij} and c_{pj} may be termed *technical coefficients*, as they specify the amounts of technically necessary inputs per unit of total output of industry j . In the interregional IO model, however, the intra-regional and interregional intermediate *input coefficients*, a_{ij}^{rr} and a_{ij}^{rs} , may no longer be called “technical coefficients” as they actually are the product of a technical IO coefficient and a not-yet-discussed IO *trade coefficient*:

$$a_{ij}^{rr} = t_{ij}^{rr} a_{ij}^{\bullet r} \text{ and } a_{ij}^{sr} = t_{ij}^{sr} a_{ij}^{\bullet r}, \text{ with } \sum_s t_{ij}^{sr} = t_{ij}^{\bullet s} = 1 \quad (45.14)$$

where an \bullet indicates a summation over the corresponding index. The intra-regional trade coefficient or *self-sufficiency ratio*, t_{ij}^{rr} , is known in the literature as the *regional purchase coefficient* (RPC, Stevens and Trainer 1980). It indicates the share in the total demand for products i by industry j in r that is supplied by the domestic industry i . Note that the RPC is equal to one minus the sum of the interregional import coefficients, that is, $t_{ij}^{rr} = 1 - \sum_{s \neq r} t_{ij}^{sr}$.

Besides interregional IO tables, the literature also distinguishes various *multiregional IO tables* (MRIOTs) and multiregional input–output (MRIO) models that are based on these data (see Oosterhaven 1984; Batten and Boyce 1986). The difference between these two accounting frameworks is that MRIOTs do not contain full interregional trade matrices for intermediate and final demand, as the IRIOT of Fig. 45.3. Instead, the basic MRIOT, schematized under Chenery/Moses in Fig. 45.5, only contains $R \times R$ columns \mathbf{h}^{rs} , each with the total of the interregional sales of the N industries i in r to all customers in s . In terms of Fig. 45.3, these trade columns thus contain the combined row sum of the *bilateral* intermediate and final demand matrices \mathbf{Z}^{rs} and \mathbf{F}^{rs} (i.e., $\mathbf{h}^{rs} = \mathbf{Z}^{rs} \mathbf{i} + \mathbf{F}^{rs} \mathbf{i}$). In addition, the basic MRIOT contains R square N -by- N matrices with the technically necessary intermediate inputs by industry j and R rectangular N -by- Q matrices with final use of products from the *worldwide* industry i by category q (not shown in Fig. 45.5). In terms of Fig. 45.3, these matrices thus contain the aggregates $\mathbf{Z}^{*s} = \sum_r \mathbf{Z}^{rs} + \mathbf{Z}^{mr}$ and $\mathbf{F}^{*s} = \sum_r \mathbf{F}^{rs} + \mathbf{F}^{mr}$.

Isard	$x = (I - B)^{-1} Tf$	$t_{ij}^{rr'} \geq 0$	
Riefler Tiebout	$x = (I - B)^{-1} Tf$ for intraregional flows $x = (I - TA)^{-1} Tf$ for interregional flows	$t_{ij}^{rr'} \geq 0$ $t_{ij}^{rr'} = t_i^{rr'}$ when $i \neq i'$ and $r \neq r'$	
Chenery Moses	$x = (I - TA)^{-1} Tf$	$t_{ij}^{rr'} \geq 0$ $t_{ij}^{rr'} = t_j^{rr'}$ when $i \neq i'$	
Leontief	$x = (I - VPA)^{-1} Tf$	$t_{ij}^{rr'} \geq 0$ $t_{ij}^{rr'} = t_i^{rr'}$ when $i \neq i'$	
Leontief Strout	$x = (I - C^{-1} DA)^{-1} C^{-1} Df$	$t_{ij}^{rr'} \geq 0$ $t_{ij}^{rr'} = t_i^{rr'}$ when $i \neq i'$ or $r \neq r'$	

Notes: The compact notation is as follows:

B is the matrix of interregional input-output coefficients;

T is the matrix of trade share coefficients;

f is the vector of final regional demands;

A is the matrix of regional technical coefficients;

V is the share vector denoting proportions of total production from each region;

P is the pooling strategy of regional demand shares;

C is the share of regional production not pooled;

D is the share of total regional demand not imported from the pool.

In general, if I is the number of sectors and R the number of regions, then the maximum number of different entries assumed in the T matrix is as follows:

Isard model : $I^2 R^2$

Riefler-Tiebout model : $IR(I+R-1)$

Chenery-Moses model : IR^2

Leontief pool models : IR

Fig. 45.5 Alternatives to the full information input–output framework (Source: Batten and Boyce (1986))

Hence, this basic multiregional IOT does contain the information to estimate the technical IO coefficients, a_{ij}^r , by means of $\mathbf{A}^r = \mathbf{Z}^r(\mathbf{x}^r)^{-1}$, but it does not contain the information on all t_{ij}^{rs} for intermediate goods, nor on all t_{iq}^{sr} for final goods. Consequently, instead of Eq. (45.14), the *multiregional IO model* (Cheney 1953; Moses 1955) uses the following behavioral assumption:

$$a_{ij}^{rr} = t_{i\bullet}^{rr} a_{ij}^{\bullet r} \text{ and } a_{ij}^{sr} = t_{i\bullet}^{sr} a_{ij}^{\bullet r}, \text{ with } \sum_s t_{i\bullet}^{sr} = 1 \quad (45.15)$$

That is, it assumes that all RPCs and all import coefficients are identical across the rows of each intermediate and each final demand block of the IRIOT of Fig. 45.3. With Eq. (45.15) substituted in the right place, the solution of the MRIO model equals (see also Fig. 45.5)

$$\mathbf{x} = (\mathbf{I} - \mathbf{T}\mathbf{A})^{-1}\mathbf{T}\mathbf{f} \quad (45.16)$$

where \mathbf{T} has the structure of the block matrix shown in Fig. 45.5 under Chenery/Moses, with on the diagonal of each block the N aggregate trade coefficients, $t_{i\bullet}^r$ or $t_{i\bullet}^s$, where \mathbf{A} is a diagonal block matrix with \mathbf{A}^{sr} on its diagonal blocks, and where \mathbf{f} is a R -block column with the stacked N -columns \mathbf{F}^{*s} \mathbf{i} .

As can be seen by comparing Eq. (45.16), the solution of the MRIO model, with Eq. (45.11a), the solution of the IRIO model, both models are able to calculate dimensionally the same employment, income, and CO₂ multipliers. Therefore, both models are able to answer the same type of income, employment, or CO₂ impact questions, be it with most likely a different degree of empirical reliability.

The initial framework proposed by Isard (1951) is often termed a *full information* IRIOT because it assumes that the flows from industry i in r to *all* industries j in s are known. The multiregional framework proposed by Chenery (1953) and Moses (1955) is often called a *limited information* MRIOT because their goal was to estimate the proportion of exports from sector i in region r that would move to region s . An intermediate variant was specified by Riefler and Tiebout (1970). They worked with full information for the important intra-regional transactions and with limited information on interregional trade (see Fig. 45.5). Leontief and Strout (1963) moved in the other direction. They further simplified the multiregional data requirements by introducing the notion of *supply and demand pools* in both the regions of origin and destination. The interregional components were estimated using a *gravity model*. Subsequent research has proposed a variety of alternative estimation techniques; perhaps the most widely used now is the *maximum entropy* formulation originally proposed by Wilson (1970).

As regards the availability of input–output data, increasingly, national accounting agencies are issuing input–output data in the form of the *supply and use table* (SUT) accounting *framework*. In this system, a supply table has industries on the left axis and commodities across the top; following across the row would provide information on the number of different commodities produced by the industry at the left. With greater disaggregation, the matrix would tend to be diagonally dominant; off-diagonal entries would indicate what are termed *secondary products*. The use table has a commodity-by-industry matrix in which each column provides information on the commodities used in production by the industry at the top of the column. Through matrix manipulation, it is possible to obtain either an industry-by-industry matrix or a commodity-by-commodity matrix (see ten Raa and Rueda-Cantuche 2003, for an overview).

The advantage of the SUT framework is its greater flexibility – one can usually provide information on a much larger set of commodities than industries. Further, sales to final demand are of commodities thus facilitating easier linkage with consumption modeling and links with commodity flow information. Oosterhaven (1984) gives an overview of both families of interregional square IO frameworks and interregional rectangular SUT frameworks and corresponding models, and Jackson (1998) discussed methods of regionalizing national SUTs.

45.5 Adding Endogenous Consumption to the Interregional Model

Although important from a methodological point of view, the above interregional extension of the single-region IO model still leaves the most important component of local final demand, namely, private household consumption, unexplained. Investments and government expenditures may be endogenized in a similar way, but note that there is a danger in endogenizing more and more components of final demand. In the extreme case, this may lead to zero exogenous final demand with infinitely large multipliers (see Oosterhaven 2000, for a conclusion of an interesting debate on this issue).

Here we only discuss how to endogenize that part of local consumption demand that can directly be tied to the value added per local industry. Figure 45.6 shows the nature of this extension. The bold lines and boxes duplicate the basic interregional IO model from Fig. 45.4. The dotted separation in the two top boxes again indicates which part of regional final demand remains exogenous and which part now becomes endogenous by adding the four dotted boxes and arrows in the lower part of Fig. 45.6. This extension of the interregional IO model results in a so-called *type II interregional IO model*.

To distinguish the *indirect* effects of the type I models from the effects through consumption demand that are added in the corresponding type II models, the latter impacts are often referred to as *induced* effects. Figure 45.6 shows the three types of induced effects that are added in type II models. First, each of the two type II single-region models adds *intra-regional induced effects* to their own intra-regional indirect effects, by making the intra-regional sales of consumption goods (c^{rr} and c^{ss}) dependent on the own region's value added, and thus on the own region's total output. Secondly, the type II interregional model adds *induced spillover effects* to the indirect spillover effects, by making the exports of consumption goods by region r and s (c^{rs} and c^{sr}), and thus the output of the exporting regions, dependent on the value added of the purchasing region s and r . Finally, combining these two induced spillover effects, and enhancing the product with the intra-regional induced effects, the interregional type II model adds *induced feedback effects* to the indirect feedback effects of type I model.

Next, we discuss the nature of the induced effects in more detail. In the type I interregional IO model, all consumption demand of households living in region r is exogenous. The same holds for the consumption of households in s . Some part of private consumption, however, will directly depend on the size and growth of the own regional value added. Some other part will only be influenced indirectly, that is, after the *interregional redistribution* of regionally generated value added through interregional commuting, interregional capital income flows, and central government's social security and taxation schemes. (The empirical specification of these relations requires the information that would be available in interregional Social Accounting Matrices (SAMs; see Pyat and Round (1979), for the original national SAM). SAMs usually distinguish between the supply of products and the use of them by industries and final demand by incorporating a SUT. Besides a SUT,

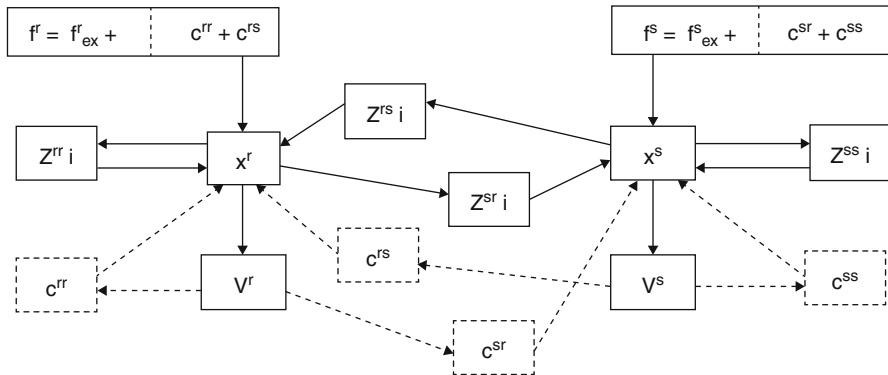


Fig. 45.6 The causal structure of the endogenous consumption extension. Legend: see Fig. 45.4. In addition: \mathbf{f}_{ex}^r , vector with remaining exogenous final demand of the type II interregional IO model; \mathbf{c}^{rs} , vector with the endogenous delivery of consumption goods by sectors i in region r to households in region s

a SAM usually contains extensive information on the generation and redistribution of all kind of income, directly and through government taxes and social security programs, mostly for a series of different types of households. There are, however, few interregional SAMs.) Here, we only add a simple direct relation between regional value added (V^r) and that part of the own region's private consumption that is directly depending on it (\mathbf{c}^{rr} and \mathbf{c}^{sr}). A comparable relation is added for the other region's private consumption (\mathbf{c}^{ss} and \mathbf{c}^{rs} ; see Fig. 45.6).

To better understand the nature of this addition, we first discuss the determination of the endogenous part of the consumption of *households living in r* in normal algebra:

$$c_{ij}^{rr} = q_{ij}^{rr} x_j^r = t_{ic}^{rr} a_{ic}^{rr} (1 - s^r) (1 - t^r) w_j^{rr} x_j^r, \text{ for all } i, j \quad (45.17a)$$

$$c_{ij}^{sr} = q_{ij}^{sr} x_j^s = t_{ic}^{sr} a_{ic}^{sr} (1 - s^r) (1 - t^r) w_j^{sr} x_j^r, \text{ for all } i, j \quad (45.17b)$$

The typical coefficient q_{ij}^{rr} of Eq. (45.17a) indicates the amount of goods and services produced by industry i in r that is consumed by households living in r and earning their income in industry j in r , per unit of output of that industry.

This *consumption demand coefficient* is built up from a series of separate coefficients. Working backward in the formula, but along the chain of cause and effect, the *labor income coefficient* w_j^{rr} indicates the gross labor income earned by households living in r per unit of output of industry j in r . The *regional tax rate* t^r determines which part of that income is disposable for consumption. The *regional savings ratio* s^r determines which part of that disposable income is actually consumed. The regional *consumption package coefficient* a_{ic}^{rr} indicate which part of total consumption in r is spent on products of the worldwide industry i .

Finally, *consumption self-sufficiency ratio* t_{ic}^{rr} indicate which part of the total consumption of products i originates from the own region r .

The consumption demand coefficient q_{ij}^{sr} of Eq. (45.17b) indicates the imports of consumption goods from industry i in s for households in r earning a labor income in industry j in r , per unit of industry j 's output. Its built-up is the same as that of q_{ij}^{rr} , except for the self-sufficiency ratio t_{ic}^{rr} that are replaced with the import coefficients t_{ic}^{sr} . The comparable consumption demand coefficient, q_{ij}^{ss} and q_{ij}^{rs} , for households living in region s , are constructed in the same way as Eq. (45.17a, b).

Combining these four sets of consumption demand coefficients, one can summarize the endogenous part of private consumption by households in matrix algebra:

$$\begin{bmatrix} \mathbf{c}^{rr} \\ \mathbf{c}^{sr} \\ \mathbf{c}^{ss} \end{bmatrix} + \begin{bmatrix} \mathbf{c}^{rs} \\ \mathbf{c}^{ss} \end{bmatrix} = \begin{bmatrix} \mathbf{q}^{rr} \\ \mathbf{q}^{sr} \end{bmatrix} \mathbf{x}^r + \begin{bmatrix} \mathbf{q}^{rs} \\ \mathbf{q}^{ss} \end{bmatrix} \mathbf{x}^s = \begin{bmatrix} \mathbf{q}^{rr} & \mathbf{q}^{rs} \\ \mathbf{q}^{sr} & \mathbf{q}^{ss} \end{bmatrix} \begin{bmatrix} \mathbf{x}^r \\ \mathbf{x}^s \end{bmatrix} = \mathbf{Q} \mathbf{x} \quad (45.18)$$

The mathematical solution of the type II interregional IO model is simple, but first, the final demand from the accounting identity of the type I model of Eq. (45.12a, b) needs to be split up into the endogenous consumption demand of Eq. (45.18) and remaining exogenous final demand \mathbf{f}^{ex} (see also Fig. 45.6):

$$\begin{bmatrix} \mathbf{x}^r \\ \mathbf{x}^s \end{bmatrix} = \begin{bmatrix} \mathbf{Z}^{rr} & \mathbf{Z}^{rs} \\ \mathbf{Z}^{sr} & \mathbf{Z}^{ss} \end{bmatrix} \begin{bmatrix} \mathbf{i} \\ \mathbf{i} \end{bmatrix} + \begin{bmatrix} \mathbf{f}^r \\ \mathbf{f}^s \end{bmatrix} = \mathbf{Z} \mathbf{i} + \begin{bmatrix} \mathbf{c}^{rr} \\ \mathbf{c}^{sr} \\ \mathbf{c}^{ss} \end{bmatrix} + \begin{bmatrix} \mathbf{f}^r \\ \mathbf{f}^s \end{bmatrix}^{\text{ex}} \quad (45.19)$$

Then, substitution of $\mathbf{Z} \mathbf{i} = \mathbf{A} \mathbf{x}$ and Eq. (45.18) in Eq. (45.19), transfer of $\mathbf{A} \mathbf{x}$ and $\mathbf{Q} \mathbf{x}$ to the left-hand side, and pre-multiplication of both sides with $(\mathbf{I} - \mathbf{A} - \mathbf{Q})^{-1}$ provide the type II model's solution for endogenous output:

$$\mathbf{x} = (\mathbf{I} - \mathbf{A} - \mathbf{Q})^{-1} \mathbf{f}^{\text{ex}} = \mathbf{L}^{**} \mathbf{f}^{\text{ex}} \quad (45.20)$$

In Eq. (45.20), \mathbf{L}^{**} represents the *type II interregional Leontief inverse*, indicating the direct, indirect, and *induced* interregional impacts of any change in the remaining exogenous demand \mathbf{f}^{ex} in region r or s . Naturally, each type II output, employment, or income multiplier is larger than its type I equivalent, but, as before, this is compensated by a smaller exogenous demand \mathbf{f}^{ex} compared to the type I exogenous demand \mathbf{f} , as shown in Eq. (45.19) and Fig. 45.6. And, consequently, endogenous output, employment, and income are the same in both models.

This means that essentially the same type of applications that are done with type I interregional IO models may also be done with a type II model. The same holds for the type II price dual that has a causal structure and interpretation that runs along arrows in directions opposite to those of the type II quantity model shown in Fig. 45.6 (as explained in Sect. 45.3). Its main difference with the type I price model is that regional wage rates in a type II price model are endogenous and not exogenous. This allows for the analysis of interregional, interindustry price/wage/price inflationary processes (cf. Oosterhaven 1981a).

45.6 Further Demo-economic Extensions of the Interregional IO Model

The size of the intra-regional type II multipliers, and especially the interregional type II spillovers, would again be larger if the interregional redistribution of labor incomes through commuting and cross-border shopping trips would be incorporated in the \mathbf{Q} matrix in Eq. (45.18). There is, in fact, a whole family of demo-economic extensions of the basic type I model into type II, III, etc., single-region IO models (see Batey 1985).

This literature makes a distinction between (i) increases in labor incomes accruing to resident workers (*intensive income growth*), (ii) new labor incomes accruing to migrants and unemployed (*extensive income growth*), (iii) and the loss of benefits of formally unemployed (*redistributive income growth*). To estimate the induced consumption effects, (i) *intensive* income growth requires the use of marginal instead of average consumption demand coefficients, (ii) *extensive* income growth of migrants can be handled with the average consumption demand coefficients as in Eq. (45.17a, b), while (iii) *redistributive* income growth of unemployed requires using the difference between the average consumption coefficients of workers and unemployed. Hence, these three types of income change can only be modeled properly if levels of economic activity are explicitly distinguished from changes therein.

With levels and changes in levels distinguished, the interregional *type III demo-economic IO model* solution for, for example, endogenous employment becomes

$$\begin{aligned}\Delta \mathbf{v} &\equiv \hat{\mathbf{c}} \Delta \mathbf{x} + \Delta \hat{\mathbf{c}} \mathbf{x}_{-1} = \hat{\mathbf{c}} (\mathbf{I} - \mathbf{A} - \mathbf{Q}_w + \mathbf{Q}_u)^{-1} \Delta \mathbf{f} + \Delta \hat{\mathbf{c}} \mathbf{x}_{-1} \\ &= \hat{\mathbf{c}} \mathbf{L}^{**} \Delta \mathbf{f} + \Delta \hat{\mathbf{c}} \mathbf{x}_{-1}\end{aligned}\quad (45.21)$$

where $\Delta \mathbf{v}$ = interregional *NR*-vector with the employment change by industry, by region, $\Delta \hat{\mathbf{c}}$ = decreases in employment coefficients due to *nominal labor productivity growth*, and \mathbf{x}_{-1} = output impact of the combined lagged endogenous and exogenous variables. Furthermore, \mathbf{Q}_w and \mathbf{Q}_u represent the *NR*-by-*NR* matrices with consumption demand coefficients, indicating the private consumption of products from industry i in r , respectively, per working resident and per unemployed resident previously working in industry j in s , per unit of output of industry j in s . Note the positive sign in front of \mathbf{Q}_u , which indicates the *negative feedback effect* of employment growth on unemployment benefits in this type III model.

Typically, \mathbf{Q}_w and \mathbf{Q}_u need to be jointly specified by an IO vacancy-chain submodel, which determines which vacancies in industry j in s are filled by workers from industry i in r and which are filled by, for example, school-leavers and unemployed. With the unemployment benefits of the Netherlands, the intra-regional type III multipliers with vacancy chains $\hat{\mathbf{c}} (\mathbf{I} - \mathbf{A} - \mathbf{Q}_w + \mathbf{Q}_u)^{-1}$ move between 35 % and 60 % of the difference between the intra-regional type I multipliers $\hat{\mathbf{c}} (\mathbf{I} - \mathbf{A})^{-1}$ and intra-regional type II multipliers $\hat{\mathbf{c}} (\mathbf{I} - \mathbf{A} - \mathbf{Q}_w)^{-1}$ per industry (van Dijk and Oosterhaven 1986, the interregional IO software package IRIOS uses a generalization of Eq. (45.20); see Stelder et al. 2000).

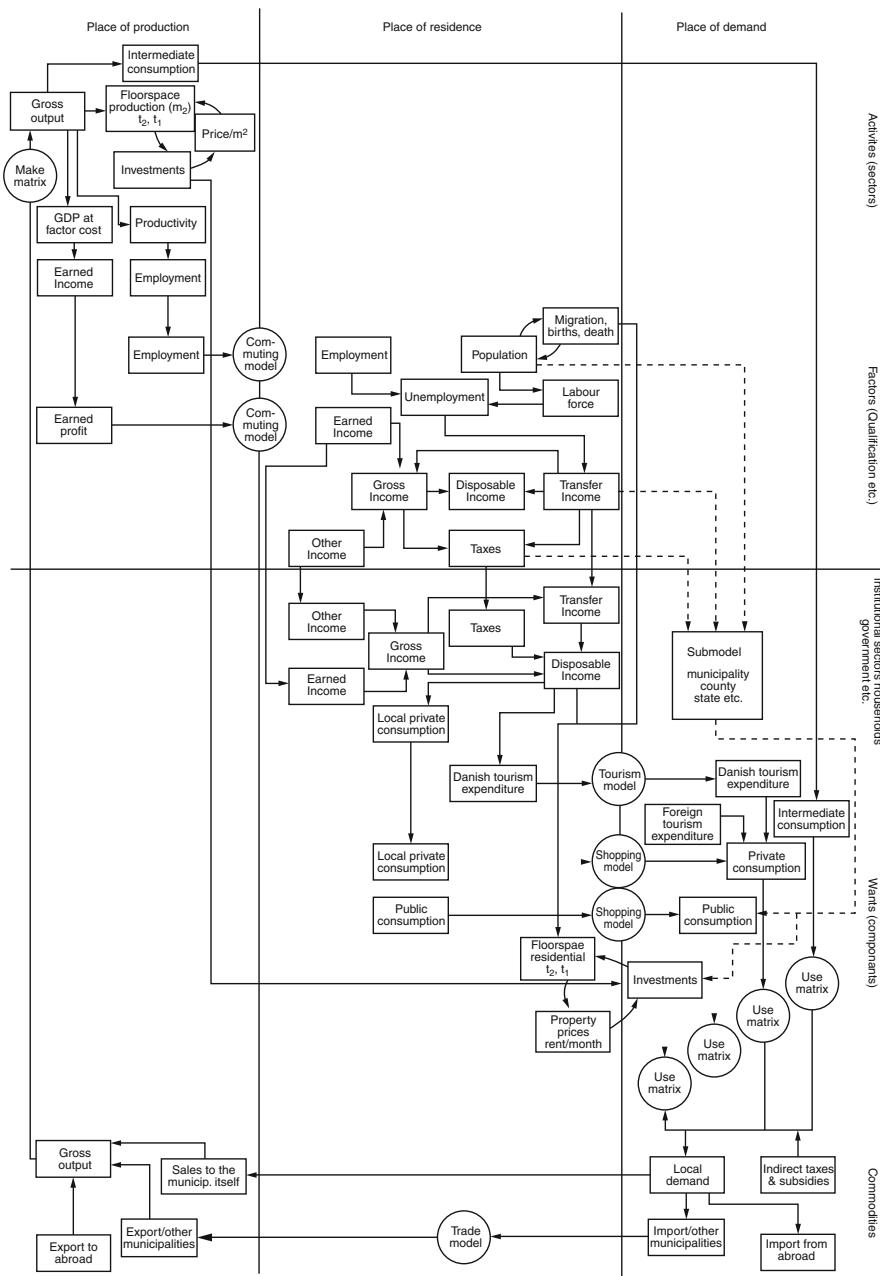


Fig. 45.7 LINE: the extended demo-economic model for the Danish economy

To finish this introductory overview of interregional input–output modeling, Fig. 45.7 shows the structure of the state-of-the-art *demo-economic model* LINE of Denmark, with 12 industries, 20 commodities, 7 ages, 2 sexes, 5 education levels for labor, 4 household consumption types, 13 private and 8 public consumption need components, 10 capital/investment components, and 277 municipalities (Madsen and Jensen-Butler 2004). Besides this data hungry detail, LINE has commuting, shopping, and domestic and foreign tourism sub-models, which are all interregional, along with taxes, social security, and central and local government budget sub-models.

Its most interesting feature, however, is not shown in Fig. 45.7. Along with the quantity model that is shown in Fig. 45.7, LINE also has a price model, with its arrows running in the opposite direction (as explained in Sect. 45.3). Finally, in contrast to the explanation of the non-interaction between prices and quantities in Sect. 45.3, LINE uses nonzero price elasticities for exogenous final demand, mainly foreign exports, and finite elasticities for the supply of exogenous primary inputs, mainly foreign imports and the impact of investments on size of the capital stock. This creates a price-quantity interaction between the quantity model and its price dual, which is mathematically solved by means of iteration. This presents a good example of the flexibility of interregional IO analysis that is able to combine an amount of detail that is only tractable in *linear* demo-economic type of models, with the theoretical advantages of the mostly much smaller, but nonlinear computable general equilibrium (CGE) models.

45.7 Conclusion

One of the new challenges, both an accounting and a modeling one, for interregional input–output analysis, will be the growth of *Internet retail purchases*. Without the possibilities of retrieving survey data, it will be difficult to allocate the flows of funds and the concomitant transfer of a physical good or service across space. In addition, whereas a retail transaction in a household’s home region may not involve the purchase of transportation costs (these would be part of the price mark up by the retailer), an Internet purchase of a good often is accompanied by mailing or delivery charges. As Internet retail sales increase in importance, tracing these flows and carefully allocating the components to appropriate locations will present new challenges for modelers.

With advances in computer software, modeling systems have become more sophisticated with the result that increasingly input–output components are nested within larger models. In many cases, the larger models represent a trajectory to make more activity endogenous and to embrace more received theory to help interpret the behavior of aggregates (e.g., household consumption) that were relegated to exogenous categories. However, there are times when a simple input–output system will suffice (e.g., the impact of a short-term event); whereas, in other cases, large multiregional CGE systems would be more appropriate, for example, to

study significant policy changes such as trade liberalization. Viewing the input–output system as part of a family of options is probably the most appropriate way to gauge its value and importance.

Perhaps, the most important challenge will be to embrace more detail in the household income and consumption components of these models. Households are becoming more mobile, intra- and internationally, they are changing in size and composition (e.g., the growth of two-earner households), and in almost all countries, the share of the population over 65 years of age is growing rapidly. These changes will generate important signals for the future composition of industry, the variety of goods and services produced, and the location of this production. As supply chains become geographically more dispersed, the sets of data contained in interregional input–output tables will become ever more valuable.

Acknowledgments The authors thank the editors, Piet Rietveld and Manfred Fischer, and Dirk Stelder for useful comments.

Appendix: The Microeconomic Foundation of the Leontief and the Ghosh IO model

The basic Leontief price and quantity model for a closed economy, introduced in Sects. 45.2 and 45.3, may be derived from microeconomics by assuming that all firms in each industry sell that industry's single homogeneous output under full competition, while they minimize their cost at given prices under a *Walras-Leontief production function*:

$$x_j = \min(z_{ij}/a_{ij}, \forall i; v_{pj}/c_{pj}, \forall p) \quad (45.22)$$

This results in a perfectly elastic supply of that single homogeneous output and a perfectly inelastic demand for intermediate and primary inputs, z_{ij} and c_{pj} , under fixed input ratios, a_{ij} and c_{pj} . Consequently, any change in the exogenous primary input prices is entirely and precisely passed on to all intermediate and final markets for the output of that firm. The left-hand side of Fig. 45.8 summarizes these individual firm assumptions and adds the assumptions about what is determined exogenously and what follows endogenously for the economy as a whole.

The interregional model extension, introduced in Sect. 45.4, adds *trade coefficients* to the basic Leontief model for a closed economy. The theoretical foundation for assuming trade coefficients to be fixed is less convincing than that for the technical coefficients by means of Eq. (45.22). It may be assumed that the output of, for example, agriculture is a different product in each different region. The trade coefficients will then have a technical character and will be fixed for the same reason. As each cell then relates to different goods, this assumption fits best with the “ideal,” full information, interregional IO model. It may also be assumed that the products of, for example, agriculture in different regions are close substitutes for

each other. The trade coefficients will then be fixed only for as long as the relative prices of agricultural outputs from different regions remain unchanged. As relative prices will influence all trade coefficients along a row of the IO table in the same manner, this assumption fits best with the limited information, multiregional IO model.

The left-hand side of Fig. 45.9 shows the implications of the above assumptions for the working of, for example, an individual intermediate input market. The vertical demand for these inputs is determined by the Leontief quantity model, whereas the horizontal supply of these inputs is determined by the Leontief price model. Any change in the demand of the purchasers is matched exactly by a corresponding change in its supply, without any change in the price asked by the suppliers. Hence, *demand drives the quantity model*. On the other side of the market, any change in the price asked by the suppliers is accepted by its purchasers, without any effect on their demand for this input. Hence, *cost pushes the price model*. Clearly, in the short run, this is not a realistic model unless there is excess capacity on all relevant primary input markets, whereas, in the long run, this model is only realistic if the *relative* prices of the primary inputs do not change.

The obvious follow-up question is whether the alternative IO quantity model of Ghosh (1958), and its dual price model (Oosterhaven 1996), offers a more plausible alternative. Ghosh developed his alternative IO model for the essentially centrally planned Indian economy of that time. Here we interpret the Ghosh model as a model for a market economy. As such it represents the pure opposite of the Leontief model, as can be seen by comparing the right- and left-hand side of Figs. 45.8 and 45.9.

In the Ghosh quantity model, the homogeneity assumption is made for all inputs along the columns of the IOT, instead of for all outputs along the rows of the IOT, as in the Leontief model. This implies that all inputs are perfect substitutes for each other. Hence, factories may run without labor, and cars may run without gasoline. Next, the Ghosh model assumes perfect complementarity of the outputs along the rows of the IO table, which is technically plausible for chemical industries, but which has to be based on a marketing desire to service all markets with the same constant market share for other industries. This is only possible if this supply of outputs is confronted with a perfectly elastic demand for them. Hence, *supply drives the Ghosh quantity model*. See the right-hand side of Fig. 45.8 for the remaining assumptions.

The mathematics of the Ghosh quantity model is far simpler than its economics, if only because it is the pure opposite of the Leontief model. Its solution reads as follows (see Oosterhaven 1996 for details):

$$\mathbf{x}' = \mathbf{v}'(\mathbf{I} - \mathbf{B})^{-1} = \mathbf{v}'\mathbf{G} \quad \text{and} \quad \mathbf{i}'\mathbf{Y} = \mathbf{v}'\mathbf{G} \mathbf{D} \quad (45.23)$$

where $\mathbf{B} = N \times N$ -matrix of intermediate *output coefficients*, $\mathbf{D} = N \times Q$ -matrix of final output coefficients, and \mathbf{G} = the so-called *Ghosh inverse*. In contrast to the Leontief inverse, which may be used as a measure of the *backward linkages* of each

Demand-driven quantity & cost-push price model:	Supply-driven quantity & demand-pull price model:
<i>For the individual firm:</i>	
- given demand for its single homogeneous output, i.e. perfect substitution among all outputs	- given supply of its single homogeneous input, i.e. perfect substitution among all inputs
- full complementarity of all inputs (fixed input ratios)	- perfect jointness of all outputs (fixed output ratios)
- cost minimization at given input prices	- revenue maximization at given output prices
- full competition, i.e. passing on of all input price changes into the single output price	- full competition, i.e. passing on of all output price changes into the single input price
<i>For the economy as a whole:</i>	
- exogenous demand for final outputs per industry	- exogenous supply of primary inputs per industry
- endogenous demand for all inputs per industry	- endogenous supply of all outputs per industry
- perfectly elastic supply of all primary inputs, i.e. exogenous primary input prices	- perfectly elastic demand for all final outputs, i.e. exogenous final output prices
- endogenous total output prices and quantities	- endogenous total input prices and quantities

Fig. 45.8 Assumptions of the basic Leontief and Ghosh models for market economies

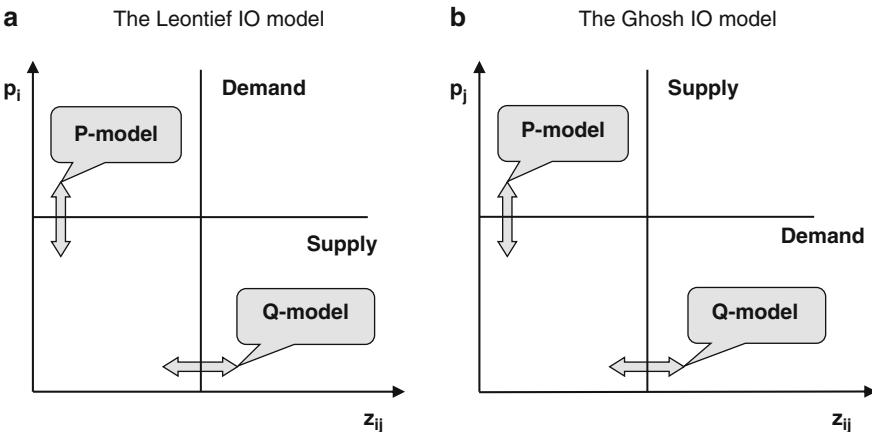


Fig. 45.9 The functioning of markets in the basic two input–output models. (a) The Leontief IO model, (b) The Ghosh IO model

sector with its direct and indirect suppliers along the columns of the IOT, the Ghosh inverse provides an indication of each sector's direct and indirect *forward linkages* with its customers, along the rows of the IOT. When used to measure forward linkages, in causal terms, the Ghosh model is best interpreted as a cost-push IO price model measured in values, instead of in prices as in Sect. 45.3 (see Dietzenbacher 1997).

The price version of the Ghosh model, which is called the *demand-pull IO price model*, is the pure opposite of the Leontief price model. Its solution reads as (see Oosterhaven 1996 for details)

$$\mathbf{p} = (\mathbf{I} - \mathbf{B})^{-1} \mathbf{D} \mathbf{p}_y = \mathbf{G} \mathbf{D} \mathbf{p}_y \quad (45.24)$$

where $\mathbf{p}_y = Q$ -vector of (index) prices for total final use per category. As opposed to the cost-push model, where \mathbf{p} refers to the price for each sector's single homogeneous output, \mathbf{p} in Eq. (45.24) relates to the price for each industry's single homogeneous *input*. Furthermore, as opposed to the cost-push model, where primary input was homogeneous across the rows, here final use is homogeneous across each column of the IO table. This assumption implies that not only firms but also consumers may drive cars without gasoline and run home appliances without electricity. See Fig. 45.8 for the remaining assumptions.

Finally, each IO market in the Ghosh model thus functions as in the right-hand side of Fig. 45.9. Prices and quantities move independently. Demand is perfectly price elastic. This means that there is infinite demand at the going market price, which is a good description of the functioning of the butter mountains and the milk lakes of the old common agricultural policy of the EU. Supply, on the other hand, is perfectly inelastic to price changes. Clearly, the Ghosh model does not offer a plausible alternative to the Leontief model, but studying it does enlarge our understanding of the nature of the Leontief model.

References

- Batey PWJ (1985) Input-output models for regional demographic-economic analysis: some structural comparisons. *Environ Plann A* 17(1):77–93
- Batten DF, Boyce DE (1986) Spatial interaction, transportation and interregional commodity flow models. In: Mills ES, Nijkamp P (eds) *Handbook in urban and regional economics*. North Holland, Amsterdam, pp 295–355
- Bouwmeester MC, Oosterhaven J, Rueda-Cantuche JM (2012). Measuring the EU value added embodied in EU foreign exports by a consolidation of 27 national supply and use tables for 2000–2007. Mimeo, University of Groningen, the Netherlands
- Chenery HB (1953) Regional analysis. In: Chenery HB, Clark PG, Vera VC (eds) *The structure and growth of the Italian economy*. U.S. Mutual Security Agency, Rome, pp 97–129
- Dietzenbacher E (1997) In vindication of the Ghosh model: a reinterpretation as a price model. *J Reg Sci* 37(4):629–651
- Ghosh A (1958) Input-output approach in an allocation system. *Econ* 25(4):58–64
- Isard W (1951) Interregional and regional input-output analysis: a model of the space economy. *Rev Econ Stat* 33(4):318–328
- Jackson RW (1998) Regionalizing national commodity-by-industry accounts. *Econ Syst Res* 10(3):223–238
- Leontief W (1936) Quantitative input and output relations in the economic system of the United States. *Rev Econ Stat* 18(3):105–125
- Leontief W (1951) *The structure of the American economy, 1919–1939*, 2nd edn. Oxford University Press, New York

- Leontief W, Strout A (1963) Multiregional input–output analysis. In: Barna T (ed) Structural interdependence and economic development. Macmillan, London, pp 119–149
- Madsen B, Jensen-Butler C (2004) Theoretical and operational issues in sub-regional economic modelling, illustrated through the development and application of the LINE model. *Econ Model* 21(3):471–508
- Miller RE, Blair PD (2009) Input–output analysis. Cambridge University Press, Cambridge, UK
- Moses LN (1955) The stability of interregional trading pattern and input–output analysis. *Am Econ Rev* 45(5):803–832
- Oosterhaven J (1981a) Export stagnation and import price inflation in an interregional input–output model. In: Buhr W, Friedrich P (eds) *Regional Development under Stagnation*. Nomos-Verlag, Baden-Baden, pp 124–148
- Oosterhaven J (1981b) Interregional input–output analysis and Dutch regional policy problems. Gower, Aldershot
- Oosterhaven J (1984) A family of square and rectangular interregional input–output tables and models. *Reg Sci Urban Econ* 14(4):565–582
- Oosterhaven J (1996) Leontief versus Ghoshian price and quantity models. *Southern Econ J* 62(3):750–759
- Oosterhaven J (2000) Lessons from the debate on Cole’s model closure. *Pap Reg Sci* 79(2):233–242
- Pyatt G, Round JI (1979) Accounting and fixed price multipliers in a social accounting matrix framework. *Econ J* 89(4):850–873
- Riefler R, Tiebout CM (1970) Interregional input–output: an empirical California-Washington model. *J Reg Sci* 10(2):135–152
- Stelder TM, Oosterhaven J, Eding GJ (2000) Interregional input–output software, IRIOS 1.0 manual. University of Groningen (downloadable at <http://www.REGroningen.nl/irios>)
- Stevens BH, Trainer GA (1980) Error generation in regional input–output analysis and its implications for nonsurvey models. In: Pleeter SP (ed) *Economic impact analysis: methodology and applications*. Martinus Nijhoff, Boston, MA, pp 68–84
- ten Raa T, Rueda-Cantuche JM (2003) The construction of input–output coefficients matrices in an axiomatic context: some further considerations. *Econ Syst Res* 15(4):441–455
- van Dijk J, Oosterhaven J (1986) Regional impacts of migrants’ expenditures: an input–output/vacancy-chain approach. In: Batey PWJ, Madden M (eds) *Integrated analysis of regional systems*. Pion, London, pp 122–147
- Wilson AG (1970) Entropy in urban and regional modelling. Pion, London

Geoffrey J. D. Hewings and Jan Oosterhaven

Contents

46.1	Introduction	903
46.2	Theories on Trade With and Without Barriers	904
46.2.1	Technological Differences: Comparative Advantage Instead of Absolute Advantage	905
46.2.2	Trade Driven by Factor Endowment Differences: The Heckscher-Ohlin Model	907
46.2.3	New Trade Theory: Economies of Scale and Love of Variety	910
46.3	Interregional Trade: Alternative Approaches	913
46.3.1	Vertical Specialization and Trade Overlap	914
46.3.2	Spatial Production Cycles	918
46.4	Interregional Trade Impacts from International Trade	920
46.5	Conclusions	922
	References	923

Abstract

Interregional trade has been relatively neglected by most trade analysts. A dearth of data has limited formal explorations of interregional trade but the magnitudes of the volumes revealed suggest that greater attention should be directed to this form of connectivity between economies. This chapter begins with a review of the theory and practice of international trade theory and its link to some of the ideas that form the basis of the New Economic Geography. Some alternative

G.J.D. Hewings (✉)

Regional Economics Applications Laboratory, University of Illinois, Urbana-Champaign,
IL, USA

e-mail: hewings@ad.uiuc.edu

J. Oosterhaven

Faculty of Economics and Business, University of Groningen, Groningen, The Netherlands
e-mail: j.oosterhaven@rug.nl

approaches to the measurement of trade are examined, especially the role of intra-industry as opposed to interindustry trade, vertical specialization, trade overlap, and spatial production cycles. Thereafter, attention is addressed to the interregional impacts of international trade.

46.1 Introduction

A press release from the Illinois state government in March 2008 announced:

Gov. Rod R. Blagojevich today announced Illinois achieved record export growth for the third consecutive year. Illinois exports totaled more than \$48.73 billion worth of goods and services in 2007, an increase of 15.79 % from 2006, according to data released from WISER, the World Institute for Strategic Economic Research, who compiles its information from the US Census Bureau, Foreign Trade Division. This record growth maintains Illinois' place as the fifth largest exporting state in the nation, up from seventh in 2005. (<http://www.illinois.gov/pressreleases>ShowPressRelease.cfm?SubjectID=3&RecNum=6691>)

Apart from the significant growth rate, the most notable feature of the news release is the absence of any mention of the growth of *interstate* or *interregional* trade. In contrast to international trade data, which are often released on a monthly or quarterly basis, data on interregional trade are often not collected at all or issued only infrequently. Hence, Gov. Blagojevich and many others have come to interpret regional trade as regional *international* trade, i.e., flows of goods and services from a region in one country to other countries, with trade between regions within the same country being ignored.

This is surprising in view of the fact that interregional trade is free of many of the restrictions imposed on international trade. Within a country, there are likely to be smaller cultural differences, lower freight costs, a uniform currency, and similar institutions. Consequently, interregional trade is most likely relatively more important than international trade. This chapter will provide analyses of the importance, structure, and measurement of interregional trade.

The next section will provide a review of international trade theory with a focus on its relevance for regional trade, i.e., with a focus on the difference between trade with and without trade barriers. The following two sections will examine some analysis of the structure of interregional trade using a variety of methods for a variety of countries. The final section provides some summary comments and challenges.

46.2 Theories on Trade With and Without Barriers

It is clarifying to start an overview of traditional trade theory by comparing it with traditional growth theory. Both are based in neoclassical economics, which means that they assume flexible prices, full competition, and substitution between inputs. Growth theory explains regional time paths of output/capita based on regional growth of factor inputs, including net in- or outflows of capital, labor, and

technology, while it assumes regional sectoral specialization to be determined exogenously. Trade theory, however, explains regional specialization from technological and factor endowment differences, and concentrates on comparative static analyses of social welfare with and without trade barriers, but it does not generate time paths of per capita output. Besides traditional trade theory, we summarize the so-called new trade theory and show how New Economic Geography is a direct descendant of it.

46.2.1 Technological Differences: Comparative Advantage Instead of Absolute Advantage

There is a host of factors that is put forward to explain the commodity pattern of interregional and international trade. David Ricardo argued in the early nineteenth century against the conventional wisdom of that time, which said that *absolute advantages* in costs determined which commodities a country could export. In fact, he showed that even countries with an absolute disadvantage in terms of the unit production cost of all its tradeable products may profitably engage in international trade without needing to protect their high cost domestic industries. He argued that even such countries must have a comparative advantage in the production of at least some goods, where *comparative advantage* is defined as the lower amount of other goods that has to be forsaken, compared to other countries, if the country at hand specializes in the production of that good.

[Figure 46.1](#) summarizes his argument in a neoclassical setting with two countries, East (E) and West (W); two products, Steel (S) and Textiles (T); one factor of production, labor (L); and constant returns to scale. [Figure 46.1](#) considers the case in which both countries are equally large (i.e., E and W have the same amount of labor available) and both have the same consumer preferences for S and T , i.e., E and W have the same *social indifference curves* (SICs), indicated by the bold convex, nonlinear lines in [Fig. 46.1](#). The falling slopes of these SICs indicate the amount of T that the consumers in East and West require to stay equally satisfied when losing one unit of S .

As there is only one factor of production that operates under constant returns to scale, the *production possibility frontiers* (PPFs) of both East and West are linear, as indicated by the bold straight lines in [Fig. 46.1](#). For each country, the slope of its PPF indicates the amount of T that the producers are able to produce more if they produce one unit less of S . The PPF of West lies entirely above that of East, which indicates that West has an absolute advantage in the production of both S and T . The PPF of East, however, has a steeper slope, which indicates that it has a comparative advantage in the production of Textiles. The equilibrium is reached where the highest SIC just touches the PPF of the country at hand.

When there is no trade between East and West, the left-hand side of [Fig. 46.1](#) shows that this set of assumptions leads to an *autarky equilibrium* with a higher level of consumption = production of both S and T in West, and thus to a higher level of welfare in West, as indicated by its higher equilibrium SIC. Also note that

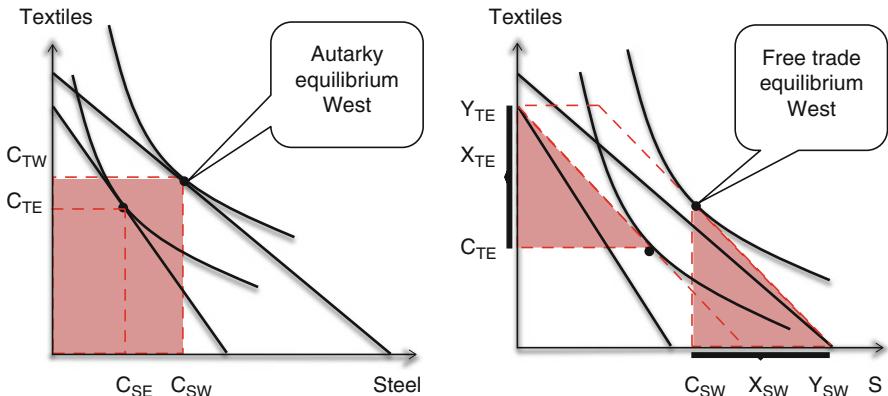


Fig. 46.1 Ricardian analysis of technological differences under autarky and free trade

the form of the shaded consumption = production rectangles of East and West indicates that the consumers in East consume relatively more Textiles (i.e., $C_{TE}/C_{SE} > C_{TW}/C_{SW}$). The explanation is that East has a comparative advantage in producing Textiles, and thus a higher domestic price of Steel to price of Textiles ratio (P_S/P_T), indicated by the higher slope of its PPF.

Removing trade barriers, including transport cost, implies that East will start to export Textiles and West will start to export Steel until the domestic P_S/P_T ratios in East and West converge to a value in between the two *autarky price ratios* shown in the left part of Fig. 46.1. This uniform *free trade equilibrium* price ratio is shown by the slope of the two parallel dashed lines in the right part of Fig. 46.1. The lower of the two dashed lines indicates the equilibrium *consumption possibility frontier* (CPF) for East, which starts at the maximum possible production of Textiles by East (Y_{TE}). Given the equilibrium P_S/P_T slope of its CPF, East will consume C_{TE} of Textiles and export the remainder of its textiles production to West (i.e., $Y_{TE} - C_{TE} = X_{TE}$), which allows East to import the amount of steel it desires at this free trade P_S/P_T ratio.

The higher of the two dashed lines indicates the equilibrium CPF of West, which starts at its maximum possible production of Steel (Y_{SW}). Given its CPF, West will consume C_{SW} of Steel and export the remainder of its steel production to East (i.e., $Y_{SW} - C_{SW} = X_{SW}$). In this two-region case, the exports of East equal the imports of West and vice versa. Hence, the two shaded *trade triangles* have exactly the same size and form. Note that this free trade equilibrium leads to a higher level of welfare for both East and West, as argued by Ricardo, but West still has a higher level of welfare as it is able to consume absolutely more Steel and more Textiles.

One of the criticisms of the Ricardo model is that it presents no explanation for the productivity differences between East and West. A more serious criticism is that it seems to predict that each country will produce and export only one product and import the remaining products that it wishes to consume with the revenue of that single export. Reality, however, shows that most countries export a whole range of

products. This criticism is not entirely correct as the production capacity of many smaller countries is limited, which reduces the import possibilities of the larger countries, as indicated by the horizontal section of the higher of the two dashed CPFs in the right part of Fig. 46.1. Consequently, larger countries need to produce more tradeable products than the single one they export. In terms of Fig. 46.1, this will lead to a corner solution for West at the kink of its consumption possibility frontier.

Note that even in a free trade situation with no interregional differences in tastes, technology, or factor endowments, the existence of transportation cost with increasing returns to transportation may result in a commodity composition of interregional trades that is opposite to those predicted by the Ricardian model (see Cukrowski and Fischer 2000). Still also in that case, additional gains from trade may emerge from the reductions of transportation costs. See Krugman et al. (2011) for a further evaluation of the Ricardian model.

46.2.2 Trade Driven by Factor Endowment Differences: The Heckscher-Ohlin Model

Above, we have formulated the Ricardian model in terms of a neoclassical trade model, although it is usually considered to be part of the classical tradition in economics (e.g., van Marrewijk 2002). The real neoclassical trade model was developed by Heckscher and Ohlin in the 1920s (hereafter called the HO model; see Leamer [1995] for a review). It also has two countries, say *E* and *W*, and two products, say *S* and *T*, which are both produced under constant returns to scale. In contrast to the Ricardian model, the HO model assumes that production technologies are identical across countries and that production requires two factors, say capital (*K*) and labor (*L*), instead of just one. Both factors of production operate under diminishing marginal returns, which means that the production possibility frontiers of both East and West are concave, nonlinear curves, as shown in Fig. 46.2, instead of straight PPFs in Fig. 46.1.

The only difference between East and West in the HO model is that they are differently endowed with *K* and *L*. Assume that say West has relatively more capital available and East relatively more labor, and assume further that producing say Textiles requires relatively more labor, while Steel requires more capital. In that case, the PPF of East has a steeper slope than that of West at all amounts of Steel being produced, as shown in Fig. 46.2. Consequently, in the *autarky equilibrium*, labor-abundant East produces and consumes relatively more labor-intensive Textiles at a higher P_S/P_T ratio, whereas West does the opposite as indicated in the left part of Fig. 46.2. (Note that the higher welfare level of West, indicated by its higher equilibrium SIC, here is coincidental. It is not a consequence of any assumption made.)

When all trade barriers, including transportation cost, are removed, the difference in the domestic price ratios induces firms in East to start exporting Textiles and those in West to start exporting Steel until the domestic P_S/P_T ratios converge to

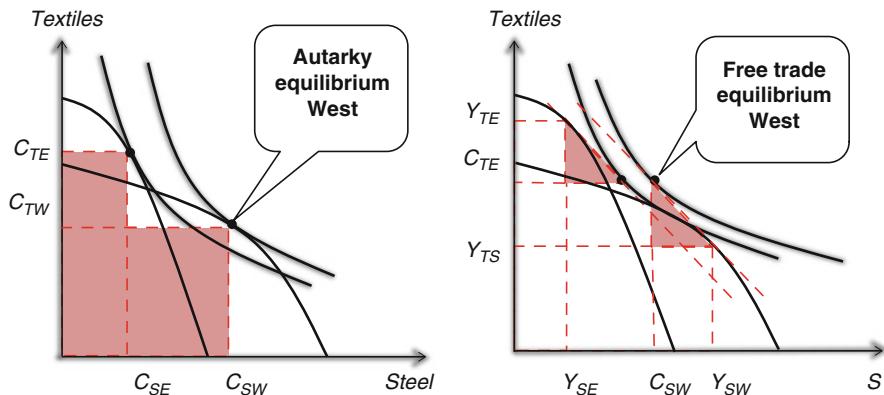


Fig. 46.2 HO analysis of factor endowment differences under autarky and free trade

a common *free trade equilibrium* price ratio. The two parallel dashed lines in the right part of Fig. 46.2 represent this equilibrium price ratio. In contrast with the Ricardian model in Fig. 46.1, the equilibrium consumption possibility frontiers (CPFs) of E and W of Fig. 46.2 do not start where the production possibility frontiers (PPFs) join the horizontal and vertical axes, but start where the PPFs equal the slope of the free trade price ratio. Consumers in East (West) consequently move away from consuming the formerly cheap Textiles (Steel) toward consuming more of the now cheaper, partly imported Steel (Textiles) until they end up at their highest indifference curve possible.

The right part of Fig. 46.2 shows that, under free trade, East exports the difference between its now larger production and smaller consumption of Textiles (i.e., $X_{TE} = Y_{TE} - C_{TE}$). Analogously, West exports the difference between its now larger production and smaller consumption of Steel (i.e., $X_{SW} = Y_{SW} - C_{SW}$; see the right part of Fig. 46.2). With these exports, E and W finance their import against the free trade price ratio of the two CPFs. This is indicated by the two shaded *trade triangles*. With only two regions, these have exactly the same size and form (like in Fig. 46.1). Again, both countries benefit from free trade, as indicated by reaching a higher social indifference curve (SIC) when moving from the left to the right part of Fig. 46.2.

In contrast to the Ricardian case, however, both countries still produce both goods, be it in different proportions than in the autarky case. This implies that both capital and labor, in both E and W , have to move from the sector that shrinks because of competing imports to the sector that grows as it becomes the exporting sector. This interindustry *production factor mobility*, of course, has consequences for the remuneration of both capital and labor (not shown in Fig. 46.2).

Under autarky, labor in the labor-abundant East will receive relatively low wages (P_L), whereas capital will receive a relatively high rate of return (P_K). The reverse will apply to West. Under free trade, relatively little labor comes free from the shrinking Steel sector of East, while relatively much is needed in its growing

Textiles sector. In contrast, relatively more capital comes free from its shrinking S , while relatively little is needed in its growing T . Consequently, in East P_L will increase, whereas P_K will decrease. As a reaction to this decrease in P_K/P_L , both the T and S sector in East will substitute away from using labor toward using relatively more capital, which partly counteracts the rising wages and declining P_K . Note that the decrease in the domestic P_K/P_L ratio in East is caused by the decrease in its domestic P_S/P_T ratio. Hence, the prices of the production factors move in the same direction as the prices of the products that use them intensively.

The reverse process in West will lead to a reverse result. Under free trade, its domestic P_S/P_T ratio will increase, which will draw both K and L from its T sector to its S sector, which will require its P_K/P_L ratio to rise too. Hence, relative product prices converge under free trade and relative factor prices follow. In fact, Paul Samuelson proved in the 1940s the *factor price equalization* (FPE) theorem: when under neoclassical conditions (i.e., identical technologies, concave PPFs, and convex SICs) both goods remain being produced in both countries, the complete equalization of product prices under free trade will lead to a complete equalization of factor returns (see van Marrewijk 2002, Chap. 5).

The Heckscher-Ohlin model of international and interregional trade is thus a full general equilibrium model that predicts sectoral production, consumption, imports and exports, and the prices of products and production factors. Its prediction of factor price equalization, however, only partly comes true in reality, whereas its prediction of the composition of exports (namely, that countries abundant with capital will export goods that use capital intensively) has been refuted many times. This outcome became known as the *Leontief paradox*, after Leontief who first used the input-output model to measure the *factor content of trade* (see Foster and Stehrer (2012) for an overview of these studies).

Leontief (1953) found that US exports embodied relatively more labor than US imports, whereas the USA was considered to be capital abundant. Trefler (1995) showed that part of the Leontief paradox may be explained by adding the Ricardian assumption of different technologies to the HO model. Others have shown that adding natural resources and different levels of human capital improves the prediction of the HO model. However, even the extended HO model still predicts far more embodied trade in the abundant factors than is found in reality. This became known as the *missing trade problem* (Trefler 1995).

Relaxing its restrictive assumptions thus improves the performance of the HO model. However, the core assumption of free trade does not hold in international trade. Even the trade between EU countries is still hampered by differences in legal systems, languages, and business cultures. The assumption of free trade, in fact, fits much better to the conditions under which *interregional trade* within one country operates. The same applies to the assumption of identical production technologies and consumer preferences, and the assumption of zero transport cost. Hence, it does not come as a surprise that extended versions of the HO model perform much better when tested on interregional trade (see Davis et al. (1997) for Japanese regions).

There is one core assumption of the HO model, however, that fits better to international than to interregional trade, namely, the immobility of factor

endowments between spatial units. The *interregional migration of production factors* K and L and the interregional mobility of products S and T have much in common. Both are motivated by price ratios (P_K/P_L and P_S/P_T) that move in the same direction when mobility barriers are removed. Both also reinforce each other's contribution to the *interregional FPE* of wages and capital returns (see Borts and Stein 1962). The HO model's prediction of sectoral specialization, however, is undermined by the interregional mobility of K and L , as it equalizes factor endowments across regions. This takes away the comparative advantage of the regions and thus undermines one of the driving forces behind what is known as *interindustry trade*, namely, why regions import and export different kinds of products.

46.2.3 New Trade Theory: Economies of Scale and Love of Variety

The Ricardian model and the HO model combined and extended, thus provide a good approach to understand interindustry trade. Both models, however, are of no help to understand why, especially, developed countries and regions import and export the same type of goods, i.e., why Germany exports as well as imports cars to and from Japan. This type of trade is known as *intra-industry trade*, and its explanation requires what is known as *new trade theory*.

The empirical importance of intra-industry trade became clear after the topical study of Grubel and Lloyd (1975). They measured the share of intra-industry trade of product i in the total trade of product i of any region r by means of the *Grubel-Lloyd index*:

$$GL_{ir} = 1 - |X_{ir} - M_{ir}| / (X_{ir} + M_{ir}) \quad (46.1)$$

where X_{ir} stands for the exports of product i by region r and M_{ir} for the imports of i by r . Brühlhart (2009) shows that the weighted average Grubel-Lloyd index for high income countries grew from 11 % in 1962 to almost 38 % in 2006, whereas it remained at a level of around 1 % for the poorest countries in that sample.

To explain this increasingly important phenomenon of intra-industry trade, two strongly related core assumptions of neoclassical economics have to be dropped, namely, that of constant returns to scale and that of full competition, while the assumption of homogeneous products has to be replaced with that of heterogeneous products and love of variety.

Introducing *increasing returns to scale* may simply be done by introducing fixed costs that are independent of the scale of production along with marginal cost (MC) that is constant per unit of output. This makes average cost (AC) a downward sloping, concave function of output, approaching MC at higher output levels, as shown in Fig. 46.3. Introducing *imperfect competition*, however, opens up a whole array of options, from pure monopoly via duopoly and oligopoly, either with or without collusion, all through to monopolistic competition. New trade theory, consequently, consists of a whole array of different models

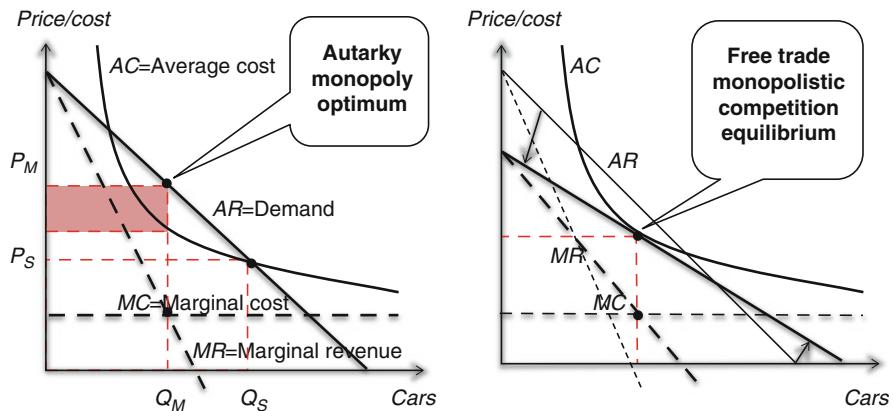


Fig. 46.3 Monopoly under autarky turning into monopolistic competition under free trade

(see Krugman et al. (2011) for an overview). We will only use the two most extreme models and thus simply move from a pure monopoly under autarky to monopolistic competition with many varieties of the same product produced under free trade in many different countries.

To start, assume a single monopoly that operates in a fully protected market, say the Trabant car company in the former German Democratic Republic (GDR). Assume that Trabant was allowed to maximize its profits, then it would have increased its sales of cars by lowering its price until its dropping marginal revenue (MR) would just equal its constant marginal cost (MC), as indicated in left part of Fig. 46.3. This would generate a maximum monopoly profit, equal to the difference between the *monopoly price* ($P_M = AR$) and average cost (AC) multiplied with the number of Trabants sold at $MR = MC$ (i.e., Q_M). This maximum profit is indicated by the shaded rectangle in left part of Fig. 46.3. In the interest of its inhabitants, however, the GDR government most likely would have ordered Trabant to further decrease its price until it just equaled average cost ($P_S = AC = AR$), at which *social price* Trabant would not have made any monopoly profits, but would have been able to produce and sell many more cars (i.e., Q_S).

Next, consider the case in which East Germany joins the European Union. Then, Trabant would have been confronted with the competition of many more car producers from many more countries. Of course, some former East Germans would have continued to buy Trabants, because that was precisely the car they wanted to have anyhow, but the majority of them would have turned to different types of cars that better suited their taste. This means dropping the assumption of a single homogenous product for the assumption of *heterogeneous products with a love of variety*. Trabant would have lost a considerable amount of domestic sales, but at the same time, it could have expanded its production by selling to new customers in the Rest of the World. In terms of market forms, the former monopolist now has to operate in multiple markets with many competitors, each producing a different variety of car, i.e., it has to operate in a market characterized by *monopolistic competition*.

In that market, the demand for Trabants is much more elastic to price changes, as indicated by the rotation of the demand curve in the right part of Fig. 46.3. In the new *free trade equilibrium*, more Trabants might have been sold, but at a lower price, which would mimic the case in which the government would force the monopoly price P_M down to AC . The monopolistic competition model further assumes *free entry and exit of firms*, which will move the demand curve for Trabants down until $AC = AR$. In reality, we saw a closing down of the Trabant car factories, probably indicating that their average cost was too high for their new much more price-sensitive demand.

This raises the question whether the welfare benefits of free trade are always positive. The most frequently used model of monopolistic competition (Dixit and Stiglitz 1977) uses a constant elasticity of substitution (CES) function in which a consumer derives utility U_c from the consumption c_i of variety i over a total of N varieties:

$$U_c = \left(\sum_{i=1}^N c_i^\rho \right)^{1/\rho} = (Nc^\rho)^{1/\rho} = \left(N^{(1/\rho)-1} \right) (Nc) \quad (46.2)$$

= love of variety resource use, $0 < \rho < 1$*

The first two terms of Eq. (46.2) represent the CES utility function, with different c_i and a *love of variety* parameter ρ . To better understand the working of this love of variety, assume that all varieties i are consumed in the same amount $c = c_i$, i.e., this consumer buys multiple TVs that all have the same size 1.0 instead of multiple TVs that have sizes of say 1.1, 1.5, and 1.9. Then, the utility U_c from consuming TVs can be decomposed as indicated by the last two terms of Eq. (46.2). They show that the increase in utility is larger than the increase in resource use due to consuming more TVs. Mathematically, this reflects that a single TV that is chosen from a large variety of TVs delivers a larger satisfaction than when there is only a single type of TV for sale.

Equation (46.2) can be used to compute the main *welfare benefits of free trade* under monopolistic competition on the demand side of such markets. Further benefits occur on the supply side due to lower AC , because of the larger production volumes under free trade (compare the left with the right part of Fig. 46.3), and because of productivity gains and innovation due to competition. The closing down of some firms, like that of Trabant, however, shows that the economic and social cost of transition may be sizeable. Besides, several cases of less usual combinations of assumptions also lead to negative impacts of free trade, such as the case of the *infant industry argument* and comparable unfortunate *path dependencies* (see van Marrewijk (2002) and Krugman et al. (2011) for further discussion).

One last, major benefit of free trade has not been discussed yet. The love of variety effect of Eq. (46.2) not only applies to final goods but also applies to *intermediate goods* and services used by firms. In that case, Eq. (46.2) mathematically reflects that, say, having to buy with one local, general purpose public relations (PR) firm delivers a less effective advertisement campaign than when

the purchasing firm may choose between a host of local specialized PR firms, normally available in big urban agglomerations. These *matching benefits of thick markets* not only apply to intermediate input markets but also to thick local labor markets. Besides these matching effects, big urban agglomerations also benefit from pecuniary external economies of scale, such as the lower risks and lower cost of *outsourcing*, and pure technical *externalities*, such as the free exchange of, especially, tacit type of information leading to more innovation.

Not surprisingly, therefore, the above monopolistic competition variant of new trade theory may be considered as the forerunner of *New Economic Geography* (NEG). In fact, simplified to its bare essentials, the first core model of NEG (Krugman 1991) only adds the mobility of labor to this variant of new trade theory (Krugman 1979).

46.3 Interregional Trade: Alternative Approaches

The above theoretical expositions have been expanded to include additional dimensions of trade, such as the links between trade and *production chains*. Hummels et al. (1998) introduced the concept of *vertical specialization* of production (see the left side of Fig. 46.4) to explain at least part of the empirical finding that economies were becoming increasingly integrated. For vertical specialization to occur, Hummels et al. (1998) postulated three conditions: (i) the good must be produced in multiple, sequential stages; (ii) two or more countries must specialize in some but not all stages; and (iii) at least one good in its various processing stages must cross an international border more than once. In essence, consider a good produced in a country for export that uses an imported component. Translated to the interregional system, vertical specialization would be similar with an imported component from region r being used in the production of a good in region s that is exported to region q .

This concept feeds into several related issues: how is it linked to outsourcing, fragmentation, hollowing out, and spatial production cycles? *Outsourcing* can accompany vertical specialization when a firm that formerly used domestic inputs decides to source them from another country. However, the firm using the now imported inputs would have to export the good to qualify as being engaged in vertical specialization. *Fragmentation of production* (see Jones and Kierzkowski 2005) is a process that might be considered a necessary but not a sufficient condition for vertical specialization to take place.

Referring to Fig. 46.5, during an era of high transportation costs, firms organized production in such a way that a larger volume of products and intermediates were often produced within the same plant or within plants located in the same vicinity. As transportation costs decreased, firms were able to exploit economies of scale by fragmenting production into more specialized components that were associated with specific geographic locations. The production chain thus spread across many economies (states or countries); if the three characteristics of vertical specialization

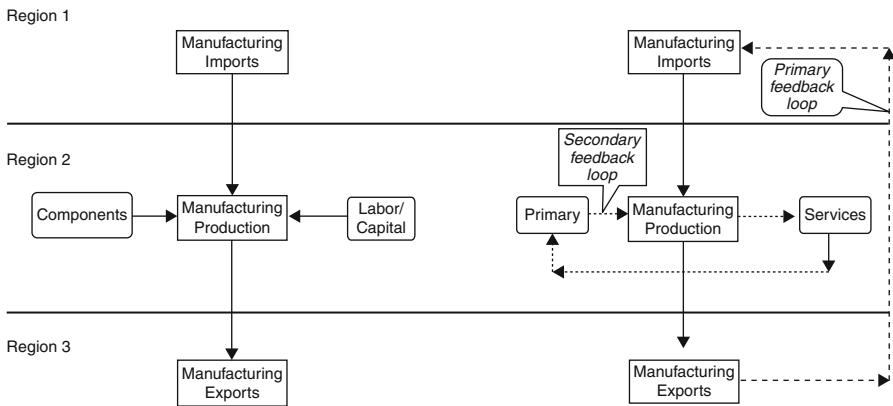


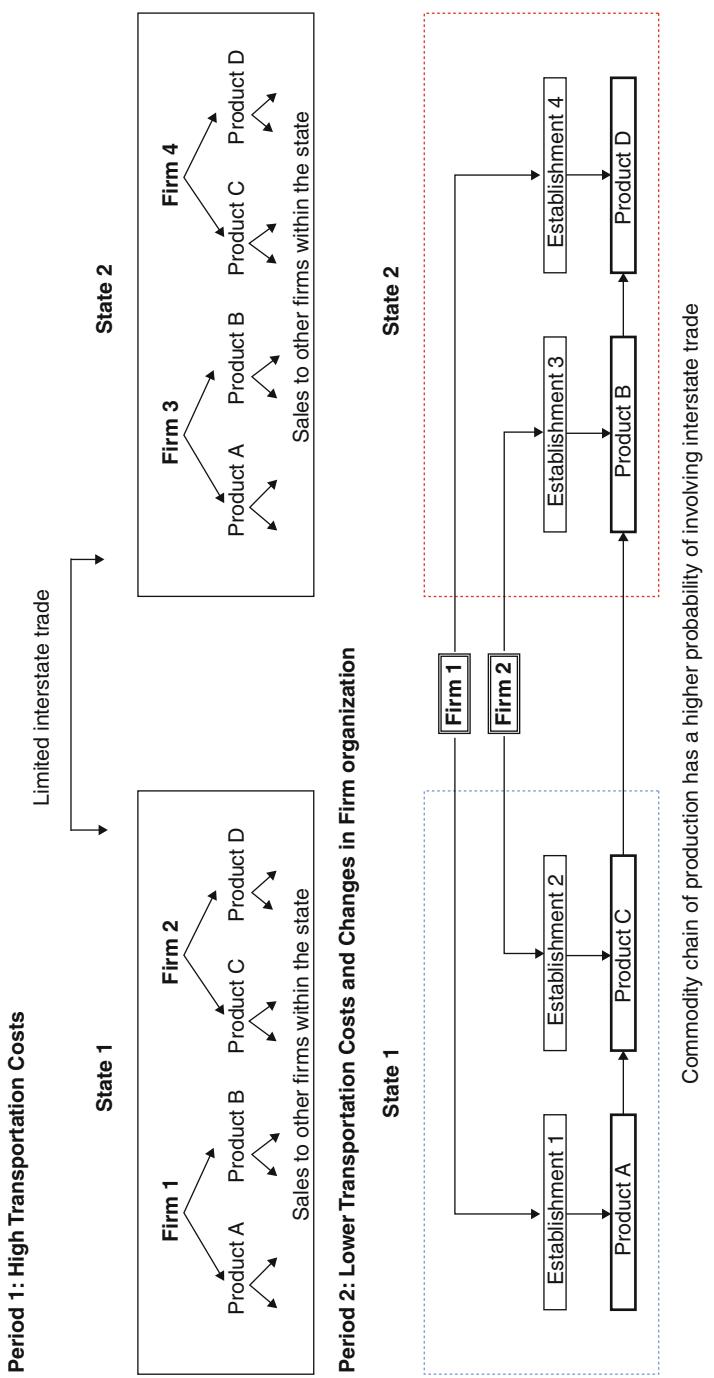
Fig. 46.4 Vertical specialization and spatial production cycles compared (Source: Sonis et al. (2002))

were met, then fragmentation would lead to vertical specialization and trade would come to be dominated by intra-industry trade.

In an economy such as Japan at the national level (Okazaki 1987) or Chicago at the regional level (Hewings et al. 1998), the process of fragmentation often accompanied the *hollowing out* of economies. This process implies that intra-economy dependence decreases and inter-economy dependence increases. The striking evidence for this are the ways in which both international and interregional trade have grown at rates exceeding domestic production. The final piece of the picture may be offered by the notion of *spatial production cycles*. Here, Sonis et al. (2002) expanded the notion of vertical specialization by exploiting the ideas of *feedback loops* (see Fig. 46.4) where the possibility that the exported good from the vertical specialization process may end up undergoing further processing until a finished good is produced that may be being exported to the country in which the whole process started.

46.3.1 Vertical Specialization and Trade Overlap

Although there have not been any attempts to measure the degree of vertical specialization in interregional trade, a companion approach by Munroe et al. (2007) attempted to measure the degree of intra-industry trade between the Midwest states of the USA. While vertical specialization focuses on the import content of exports, an important subset of this trade would be accounted for by flows between firms in the same broad industrial category. In the traditional HO model of international trade described in Sect. 46.2, trade is driven by differing factor endowments between regions. The HO model cannot adequately explain the large degree of trade in similar goods taking place among similar economies. If intra-industry trade (hereafter, IIT) is at odds with the more traditional HO framework of comparative advantage, one must first grapple with the determinants of such trade.



Commodity chain of production has a higher probability of involving interstate trade

Fig. 46.5 Changing spatial structure of production (Source: Hewings and Parr (2009))

Stone (1997) separates the determinants of IIT into two categories: industry-based determinants and regional characteristics. The industry-based determinants include product differentiation, scale economies, industry specific cost structures, and transportation costs. Regional determinants include macroeconomic characteristics, such as income level and relative capital/labor ratios. It has been assumed that IIT will increase as income differences decrease because demand structures become more similar, with fewer differences in factor endowments and growing average incomes.

Within the IIT theoretical literature, there are differing assumptions regarding the type of product differentiation within an industry that leads to IIT. The three general types of differentiation include *horizontal product differentiation* (differences of variety), *vertical product differentiation* (differences of quality), and the *vertical integration of production process itself* (trade in intermediate goods). Krugman (1991) has championed the case for horizontal differentiation leading to increased IIT, using Eq. (46.2), the Dixit-Stiglitz equation; his contributions add the fact that consumer preferences become more diverse leading to greater product differentiation by type or variety. As each region specializes in a certain variety of a good, incentives for trade arise (see Fig. 46.5). This model is most applicable to the study of trade among highly developed economies, with a predominance of trade in capital-intensive goods and a high level of technology. Intra-industry trade between economies with dissimilar endowments and levels of technology, however, is most likely the consequence of the international fragmentation of value chains.

Thom and McDowell (1999) argued that intra-industry trade takes two forms: horizontal and vertical. *Horizontal intra-industry trade* is associated with economies of scale and occurs when products are differentiated and consumers express preferences for product variety, as in the Dixit-Stiglitz formula. *Vertical intra-industry trade*, on the other hand, is similar to interindustry trade in that it exploits comparative advantage and specialization, not between different industries as with interindustry trade but within the same industry as trade in different parts and components. Price (2001) noted two trends in the fragmentation process: trends in the spatial dimension associated with economies becoming more global (in part reflected by the hollowing out phenomenon noted earlier) and trends in the specialization dimension where firms (and particularly plants) are becoming more specialized because of the enlarged market created by global demands.

To provide a brief empirical illustration, an examination of US Midwest interregional trade was conducted using the *Grubel-Lloyd* (GL) IIT index, see Eq. (46.1), to measure the amount of *trade overlap*. A value of one would imply perfect trade overlap, and a value of zero would imply perfect specialization. Comparing GL indices for the five Midwestern states is a good point of departure for understanding trade flows within this region. Table 46.1 summarizes these findings. For each of the five states, five industries with the highest (trade overlap) and lowest (trade driven by industry specialization) GL indices are reported. In addition, the state of destination is reported, where RUS stands for “Rest of the United States.”

Table 46.1 Highest and lowest Grubel-Lloyd indices for the Midwest US states^a

			State of destination		SIC	State of destination
	Most overlap	SIC		Most specialization	SIC	
Illinois	Farm products	01	Indiana	Fresh fish	09	Indiana
	Lumber and wood prods	24	Indiana	Coal	11	RUS
	Clay, concrete, glass, or stone	32	RUS ^a	Ordinance or accessories	19	RUS
	Fabricated metal products	34	Indiana	Petroleum or coal	29	RUS
	Machinery	35	Indiana	Clay, concrete, glass, or stone	32	RUS
Indiana	Farm products	01	Illinois	Fresh fish	09	Illinois
	Nonmetallic minerals	14	Ohio	Leather or leather products	31	Illinois
	Food or kindred products	20	RUS	Textile mill products	22	Ohio
	Clay, concrete, glass, or stone	32	Illinois	Furniture or fixtures	25	Ohio
	Photographic, optical instruments	38	Ohio	Coal	11	Illinois
Michigan	Machinery excluding electrical	35	Ohio	Textile mill products	22	Ohio
	Food or kindred products	20	RUS	Apparel or finished textiles	33	Illinois
	Leather or leather products	31	Ohio	Nonmetallic minerals	14	Indiana
	Primary metal products	33	RUS	Electrical machinery	36	Illinois
	Fabricated metal products	34	Ohio	Photographic, optical instruments	38	Indiana
Ohio	Nonmetallic minerals	14	Indiana	Metallic ores	10	RUS
	Rubber or miscellaneous plastic	30	Wisconsin	Ordinances or accessories	19	RUS
	Transportation equipment	37	Illinois	Apparel or other finished textiles	23	Wisconsin
	Fabricated metal products	34	Indiana	Waste or scrap materials	40	RUS
	Machinery excluding electrical	35	Michigan	Misc. freight equipment	41	RUS
Wisconsin	Rubber or misc. plastic products	30	Ohio	Farm products	01	Ohio
	Primary metal products	33	RUS	Ordinance or accessories	19	RUS
	Fabricated metal products	34	Indiana	Pulp, paper, or allied products	26	Michigan

(continued)

Table 46.1 (continued)

Most overlap	SIC	State of destination	Most specialization	SIC	State of destination
Electrical machinery equipment	36	Indiana	Leather or leather products	31	RUS
Photographic, optical instr.	38	Illinois	Misc. freight equipment	41	Illinois

^aRUS = “Rest of the USA” (Source: Munroe et al. (2007))

As predicted by new trade theory, some of the more “high-tech” industries appear in the column with the highest trade overlap, e.g., fabricated metal, transportation equipment, machinery, and food and kindred products (agricultural processing). Conversely, in the column reporting the most specialized trade, some industries appear that are more natural resource based, or have lower levels of high-tech production methods, e.g., coal, textile mill products, pulp and paper products, metallic ores, and furniture and fixtures. However, these results are somewhat equivocal. In a few cases, an industry that exhibits a high level of trade overlap for one state is specialized in another state, e.g., photographic and optical instruments, leather and leather products, and clay, concrete glass, and stone. This finding perhaps points to the complexity of these trade flows; possibly, trade driven by both intra-industry specialization and competitive advantage occurs.

Another interesting finding is that for all states, most of the IIT is directed to other states in the Midwest. For Illinois, Ohio, and Wisconsin, more of their trade to the Midwest is driven by IIT, while their trade to states outside the Midwest is predominantly specialized. This observation underscores the importance and interdependence of trade flows among states within this region and further suggests that *agglomeration effects* are being manifested at the multistate level rather than for individual metropolitan or state economies.

It should be noted that several authors have addressed problems with the GL index. Nilsson (1997) presented two major problems with the measurement of IIT. The first is the inappropriate grouping of industry activities. He proposed an alternative measure, indicating that the volume of intra-industry trade between two countries r and s may be divided with the total number of products they trade with each other to yield a measure of the average level of intra-industry trade per product group. Further, a dynamic GL index was suggested by Brülhart (2009) based on the concept of marginal IIT to address the problem of changes in the trade flows.

46.3.2 Spatial Production Cycles

The notion of *spatial production cycles* can be considered as a reworking of the ideas of feedback loops into a form that is compatible with the vertical specialization of production proposed by Hummels et al. (1998). Further details may be found

Table 46.2 Midwest interregional flows (1992 million US dollars)^a

	Illinois	Indiana	Michigan	Ohio	Wisconsin	RUS
Illinois	154,926	5,042	7,262	3,550	8,828	111,398
Indiana	5,798	61,858	9,220	5,271	2,240	44,317
Michigan	6,190	5,910	104,122	11,158	4,520	90,265
Ohio	3,746	4,647	20,334	139,912	2,172	77,815
Wisconsin	13,688	2,768	9,492	3,819	30,951	90,257
RUS	76,202	34,994	83,228	60,998	69,836	2,581,622

^aRUS refers to the Rest of the USA (Source: Sonis et al. (2002))

Table 46.3 Two largest spatial production cycles^a

	Illinois	Indiana	Michigan	Ohio	Wisconsin	RUS
Illinois					2	1
Indiana				2	1	
Michigan				1		2
Ohio		1	2			
Wisconsin	1	2				
RUS	2		1			

1	First Production Cycle: (RUS, Michigan, Ohio, Indiana, Wisconsin, Illinois)	25.9%
2	Second Production Cycle: (RUS, Illinois, Wisconsin, Indiana, Ohio, Michigan)	23.3%

^aRUS refers to the Rest of the US. Source: Sonis et al. (2002).

in Sonis et al. (2002); the graphical structure is presented in the right-hand side of Fig. 46.4. The methodology exploits the properties of *block-permutation matrices* that enable the identification of hierarchies of spatial production cycles. For the Midwest US application, the analysis was conducted at three levels: at an aggregated level in which all sectors were collapsed into one, at the level of three sectors (primary, secondary, and tertiary), and at a six-sector level into which the previous three sectors were each divided into two.

Table 46.2 shows the geographical division of the trade between the Midwest states in 1992. From this table, it is easy to calculate that the global intensity of trade in the Midwest in 1992 reached \$3.9 trillion, while the interregional trade was \$894.9 million, which is 22.2 % of all US trade. 85.5 % of Midwest interregional trade includes export and import with the Rest of the USA; the remaining 15 % that flows among the Midwest states amounts to \$135 billion. (If flows to final demand accounts were included, the total Midwest interstate flow would be of the order of \$350–\$400 billion.) Without a detailed analysis of this table, it would be difficult to identify and interpret the dominant interregional and interactivity linkages.

This is accomplished in the following tables. On the most aggregated regional level of analysis, there is the decomposition of the Midwest trade flows into five feedback loops hierarchically ordered according to the intensity (sum of flows) of trade through this loop. Table 46.3 presents the two largest feedback loops connecting all the states of Midwest and the Rest of the USA. The larger of the

two includes 25.9 % of all interregional trade. It includes the largest export flow, from Illinois to the Rest of the USA, and the largest import flow, from the Rest of the USA to Michigan. The second of these two feedback loops accounts for 23.3 % of all interregional trade and includes the largest inner Midwest trade flow, from Ohio to Michigan. It is important to note that the spatial structure of these two loops is topologically identical; they differ only in the direction of flows. This means that the Midwest economy is well developed and bilaterally balanced: to each flow, there corresponds the equivalent counterflow.

These two feedback loops together account for 49.2 % of all Midwest interregional trade. They characterize the multilateral trade connections between all Midwest states. Some further amplification can be provided; these two multilateral feedback loops together can be presented with the help of another pair of feedback loops including only bilateral trade connections. A more detailed analysis of the structure of industry-based spatial production cycles can be found in Sonis et al. (2002).

46.4 Interregional Trade Impacts from International Trade

Finally, the relation between international and interregional trade is important. The promotion of first US-Canada Free Trade Agreement (FTA), and then NAFTA, was based on the premise that an enlarged market would provide mutual benefits to participant countries. Almost all the analysis and the presentation of the outcomes were considered at the national level, but what of the spatial impacts? Using a multiregional computable general equilibrium model, Gazel et al. (1996) estimated the regional (state-level) impacts of the US-Canada FTA to be of the order of 1–2 % in the Midwest states (which had the greatest volume of trade with Canada) and up to 5 % in states like Texas with more modest levels of trade. The analysis revealed that the relative regional gains from the FTA depend on factors other than export and import share of each region with Canada and their respective economic size. As Gazel et al. (1996) noted, the internal economic structure and the nature and volume of interregional trade played an important role in determining the outcome of the regional gains from trade.

The spatial impacts of NAFTA turn out to be much more complex; one major structural change generated by this trade agreement was the significant spatial restructuring of the supply chains of many automobile companies. As a result, the sectoral impacts were often much more varied than the spatial impacts. Andresen (2009) measured the impacts of NAFTA on Canadian provinces and found that the impact on interprovincial trade was more important than province-US trade; once again, the larger impact was on within-country trade. The US results were mixed; model specifications often fail to capture complex interplay between national and interregional trade, assuming somehow that the two are not connected.

Further, the impact of international trade changes on interregional trade is often significant and spatially concentrated (Hewings and Parr 2009). Table 46.4 provides an analysis of interregional trade focusing on the Midwest and the Rest of the USA for three selected years. First, the proportion of intraregional flows (those

Table 46.4 Analysis of interregional trade in the USA, 1980–2000^a

	1980	1990	2000
Total flow	4,688,314	4,964,328	5,933,438
Intraregional flow	83.2 %	82.4 %	80.8 %
Intra-activity	31.0 %	35.5 %	37.5 %
Interactivity	52.2 %	46.9 %	43.3 %
Interregional flow	16.8 %	17.6	19.2 %
Intra-activity	7.5 %	8.5 %	10.0 %
Interactivity	9.3 %	9.1 %	9.2 %
MW and RUS flows			
MW-to-MW	13.7 %	15.0 %	17.3 %
MW-to-RUS	8.2 %	8.4 %	8.8 %
RUS-to-MW	6.1 %	6.5 %	7.0 %
RUS-to-RUS	72.0 %	71.1 %	66.8 %

^aMW Midwest states of the USA, RUS Rest of the USA (Source: Author calculations based on the US Commodity Flow Survey, Bureau of Transportation Statistics and Midwest-Rest of the US econometric input–output model developed by the Regional Economics Applications Laboratory)

Table 46.5 Indirect interregional impacts of changes in international trade: the US Midwest

	IL	IN	MI	OH	WI	Rest of Midwest	Rest US
IL	43.8	5.1	5	4.1	5.8	20	36.2
IN	5.7	42.7	8.7	7.7	3.2	19.6	32.1
MI	6.1	7.8	30.9	16.2	4.9	28.9	34.2
OH	3.9	4.6	7.6	51.9	2.6	14.8	29.5
WI	11.3	4.4	7.4	5.4	19.7	17.2	51.9
Rest US	6.4	3.5	6.7	5.8	4.1		73.5
Inter-Avg	6.7	5.1	7.1	7.8	4.1		36.8

Note: IL Illinois, IN Indiana, MI Michigan, WI Wisconsin, OH Ohio
(Source: Same as for [Table 46.4](#))

circulating within the Midwest or the Rest of the USA) declined over the two decades of the analysis, even while total flows increased. Secondly, intra-activity flows (those between the same sectors) increased while interactivity flows (between different sectors) decreased. Interregional flows accounted for an increasing share of total flows with, once again, intra-activity flows increasing and interactivity flows decreasing. Flows within the Midwest but between different states increased as did trade between the Midwest and the Rest of the USA (in both directions).

[Table 46.5](#) provides assessment of the way in which changes in international trade differentially impact regions. Even though the macrostructures of the states of Indiana (IN), Michigan (MI), Ohio (OH), and Wisconsin (WI) are similar, a change

in international demand will generate different impacts – both internally and externally. Wisconsin is far more open – only 20 % of the indirect effects of a unit change in international trade from this state remain within the state, 17.2 % leaks out to the other Midwest states, and the remainder (51.9 %) to the Rest of the USA. In contrast, Michigan retains about 30.9 % of the indirect effects, but almost an equal percentage (28.9 %) spills over to the other Midwest states, and only 34 % finds its way to the Rest of the USA. Ohio is the least “generous” with other states, retaining over 50 % of the indirect effects within its borders. The strength of these interstate connections in the Midwest – in 1993 over 40 % of each constituent state’s imports and exports were derived from or destined for other Midwest states – means that changes in international trade will have a significantly concentrated effect.

A similar analysis with Spanish regions (Llano et al. 2010) revealed that while domestic (intra- and interregional) trade flows dominated, international imports and exports grew at much faster rates between 1995 and 2005, a period within which Spain became increasingly integrated into the European Union. There is a continuing debate about the related so-called *border effect* in dampening trade flows. When considering intranational trade flows, the question has been posed as to whether state borders have an impact that is comparable to that of national borders.

Hillberry and Hummels (2003, 2008) have explored some aspects of this problem. Taking as a challenge that state borders apparently impeded trade flows, they revealed that much of the apparent limitation on interstate flows could be explained by the dominance of wholesaling activity which, by its very nature, was focused on local markets (Hillberry and Hummels 2003). With greater access to individual establishment-level data, a more extensive analysis was conducted to examine the degree to which trade responded to geographic frictions (Hillberry and Hummels 2008). In addition, they complemented the earlier work by decomposing trade into *extensive* margins (the number of commodities involved) as opposed to *intensive* margins (the value per commodity). Having access to actual trade flows (with precise distances up to a tolerance of four miles) from individual establishments, they were able to show that spatial frictions reduce the extensive margins and that the so-called home bias was an artifact of geographic aggregation.

Among other findings, they found that value declined very rapidly with distance, “...dropping off almost an entire order of magnitude between 1 and 200 miles, and is nearly flat thereafter” (Hillberry and Hummels 2008, p. 533). Further, the number of unique shipments drops at about the same rate as value over distance, but value per shipment had no clear decline with distance. In essence, they conclude that spatial frictions have their greatest impact on the number of shipments rather than on the value per shipment. Shipments within a 5-digit zip code (about a four mile radius of the shipper) are three times higher than those outside the zip code; if the results had been estimated at a 3-digit (more extensive spatial unit) level, then intrastate flows dominate.

However, it is not clear how these spatial frictions manifest themselves since the nature of state barriers vis-a-vis national ones are several orders less intrusive. In addition, the limitations in the number of goods exported/imported may be attributed to the lack of demand and thus to variations in economic structure. One further

interesting finding from their work is that even at the five-digit level, intra-industry trade occurs – further verifying that the Jones and Kierzkowski (2005) ideas even operate at very small spatial scales.

46.5 Conclusions

The analysis presented in this chapter has drawn from research that has examined gross trade flows. Increasingly, research is now focusing on *trade in value-added*. For example, the completion of the World Input–output Database (www.wiod.org) in 2012 enabled analysis that revealed that EU's trade deficit with China was 36 % smaller when the value of the separate stages of production was summed independently rather than focusing on the value of the end products shipped. Applications of such methodology at the regional level would provide the potential for some reconsideration of the nature of trade flows.

Further, the new approach might rekindle interest in Thirlwall's (1980) proposition that regional problems are balance of payments problems, an issue recently reexplored by Ramos (2007). In addition, a related avenue of exploration is the degree to which trade in people (migration) and trade in goods and services are linked. Several studies have been directed at the impact that immigrants might have on opening markets between their current and former countries; in this sense, the role of interregional trade may also play a synergistic role in *interregional migration* flows.

Earlier analysis of regional policies promoting greater diversification of state and local economies, as opposed to exploiting existing competitive advantages, now have to be extended to the portfolio of export and import dependencies. Traditional *cluster-based development* strategies are now being challenged by the increasing hollowing out of regional economies and the continuing fragmentation of production. Simple dyadic trade exchange has been replaced by complex flows; the final origin for an import or the final destination of an export from any given region may hide the chains of interaction that contributed to the assembly of the import, and the ultimate destination of the export may be many further product transformations away and these transformations may occur in more than one location. Unraveling these complexities will require detailed databases and careful integration with other sources of information. While input–output and trade tables provide information on flows between sectors and countries, they reveal little about the ordering or sequencing of trade flows. Issues of risk and vulnerability will come to assume a more critical role as notions of dependency and interdependency are further elaborated and modified to account for much more extensive considerations of trade and its role in economic growth and development.

References

- Andresen MA (2009) The geographical effects of the NAFTA on Canadian provinces. *Ann Reg Sci* 43(1):251–265

- Borts GH, Stein JL (1962) Economic growth in a free market. Columbia University Press, New York
- Brülhart M (2009) An account of global intra-industry trade, 1962–2006. *World Econ* 32(3):401–459
- Cukrowski J, Fischer MM (2000) Theory of comparative advantage: do transportation costs matter? *J Reg Sci* 40(2):311–322
- Foster N, Stehrer R (2012) The factor content of trade, a survey of the literature. *World Input–output database*, Deliverable 8.1, WIIW, Vienna
- Davis DR, Weinstein DE, Bradford SC, Shimpo K (1997) Using international and Japanese regional data to determine when the factor abundance theory of trade works. *Am Econ Rev* 87(3):421–446
- Dixit A, Stiglitz J (1977) Monopolistic competition and optimal product diversity. *Am Econ Rev* 67(3):297–308
- Gazel R, Hewings GJD, Sonis M (1996) Trade, sensitivity and feedbacks: interregional impacts of the US–Canada free trade agreement. In: van den Bergh JCJM, Nijkamp P, Rietveld P (eds) *Recent advances in spatial equilibrium modeling*. Springer, New York/Berlin/Heidelberg, pp 278–300
- Grubel HG, Lloyd PJ (1975) *Intra-industry trade: the theory and measurement of international trade in differentiated products*. Wiley, New York
- Hewings GJD, Sonis M, Guo J, Israilevich PR, Schindler GR (1998) The hollowing out process in the Chicago economy, 1975–2015. *Geogr An* 30(3):217–233
- Hewings GJD, Parr JB (2009) The changing structure of trade and interdependence in a mature economy: the US Midwest. In: McCann P (ed) *Technological change and mature industrial regions: firms, knowledge, and policy*. Edward Elgar, Cheltenham, pp 64–84
- Hillberry R, Hummels D (2003) Intranational home bias: some explanations. *Rev Econ Stat* 85(4):1089–1092
- Hillberry R, Hummels D (2008) Trade responses to geographic frictions: a decomposition using micro data. *Eur Econ Rev* 52(3):527–550
- Hummels D, Rapoport D, Yi KM (1998) Vertical specialization and the changing nature of world trade. *Econ Policy Rev*, Federal Reserve Bank of New York, June 1998, pp 79–99
- Jones RW, Kierzkowski H (2005) International fragmentation and the new economic geography. *N Am J Econ Financ* 16(1):1–10
- Krugman PR (1979) Increasing returns, monopolistic competition, and international trade. *J Int Econ* 9(4):469–479
- Krugman PR (1991) Increasing returns and economic geography. *J Political Econ* 99(3):483–499
- Krugman PR, Obstfeld M, Melitz M (2011) *International economics, theory & policy*. Addison Wesley, Boston
- Leamer EE (1995) The Heckscher-Ohlin model in theory and practice, vol 77, Princeton studies in international finance. Princeton University Press, Princeton
- Leontief WW (1953) Domestic production and foreign trade: the American capital position re-examined. *Proc Am Philos Soc* 97(4):331–349
- Llano C, Esteban A, Perez J, Pulido A (2010) Opening the interregional trade “black box:” the C-intereg database for the spanish economy (1995–2005). *Int Reg Sci Rev* 33(3): 302–337
- van Marrewijk C (2002) *International trade & the world economy*. Oxford University Press, Oxford
- Munroe DK, Hewings GJD, Guo D (2007) The role of intraindustry trade in interregional trade in the Midwest of the US. In: Cooper RJ, Donaghy KP, Hewings GJD (eds) *Globalization and regional economic modeling*. Springer, New York/Berlin/Heidelberg, pp 87–105
- Nilsson L (1997) The measurement of intraindustry trade between unequal partners. *Weltwirtschaftliches Arch* 133(3):554–565
- Okazaki F (1987) General verification of the logit-type stochastic trade pattern using intertemporal, interregional input–output data. *Papers Reg Sci Assoc* 63(1):1–11

- Price VC (2001) Some causes and consequences of fragmentation. In: Arndt SW, Kierzkowski H (eds) *Fragmentation: new production patterns in the world economy*. Oxford University Press, New York, pp 88–107
- Ramos PN (2007) Does the trade balance really matter for regions? *Ann Reg Sci* 41(1):229–243
- Sonis M, Hewings GJD, Okuyama Y (2002) Vertical specialization and spatial production cycles in interregional trade: feedback loops analysis of the Midwest economy. In: Hewings GJD, Sonis M, Boyce D (eds) *Trade, networks and hierarchies, advances in spatial sciences*. Springer, New York/Berlin/Heidelberg, pp 347–364
- Stone LL (1997) The growth of intraindustry trade. Garland Publishing, New York
- Thirlwall A (1980) Regional problems are “balance-of-payments” problems. *Reg Stud* 14(5):419–425
- Thom R, McDowell M (1999) Measuring marginal intraindustry trade. *Weltwirtschaftliches Arch* 135(1):48–61
- Trefler D (1995) The case of missing trade and other HOV mysteries. *Am Econ Rev* 85(5):1029–1046

Section VI

Environmental and Natural Resources

environmental

world future site production
public space goods
equilibrium
natural price labor
countries economic risk travel
resources property water
demand environment agricultural

impacts resource

Yacov Tsur and Amos Zemel

Contents

47.1	Introduction	930
47.2	The Canonical Resource Management Model	931
47.3	Resource Management Under Uncertainty	933
47.3.1	Uncertain T	933
47.3.2	Stochastic Stock Dynamics	938
47.3.3	Discounting	939
47.3.4	Instantaneous Benefit	940
47.3.5	Post-planning Value	940
47.3.6	Compound Uncertainties	941
47.4	Integrating Natural Resources and Aggregate Growth Models	943
47.4.1	An Integrated Model	943
47.4.2	Uncertainty in the Integrated Model	944
47.5	Irreversibility and Uncertainty	945
47.6	Knightian Uncertainty	946
47.7	Conclusions	946
	References	947

Abstract

Uncertainty affects the dynamic trade-offs of environmental and natural resource management in a variety of ways and forms. The uncertain responses to anthropogenic activities may be due to genuine stochastic processes that drive

Y. Tsur (✉)

Department of Agricultural Economics and Management, The Hebrew University of Jerusalem,
Rehovot, Israel

e-mail: yacov.tsur@mail.huji.ac.il

A. Zemel

Department of Solar Energy and Environmental Physics, Jacob Blaustein Institutes for Desert
Research, Ben Gurion University of the Negev, Sede Boker Campus, Israel
e-mail: amos@bgu.ac.il

the evolution of the underlying natural systems or simply due to our poor understanding of these complex systems and their interactions with the exploitation policies. These interactions are of particular importance when the ecosystem response might involve irreversibility, so that unexpected undesirable outcomes cannot be undone after they are realized. In this chapter, we review the various sources of uncertainty, the methodologies developed to account for them, and the implications regarding the management of environmental and natural resources.

47.1 Introduction

Environmental and resource economics is the branch of economics in which human activities interact with natural processes, giving rise to complex dynamical systems. Since the natural processes that constrain the options open to resource managers evolve in ways that are often poorly understood, the responsible management of natural resources must account for the dynamical and uncertainty aspects of the combined human-natural systems. These two aspects make the central theme of this chapter.

The importance of uncertainty considerations in the design of environmental policies has long been recognized, and the literature dealing with this topic is vast (see Mangel (1985) and the recent reviews of Heal and Kriström (2002) and Pindyck (2007)). In this chapter, we consider this issue emphasizing the rich variety of forms in which uncertainty enters all components of the management problems. Uncertainty stems from two main sources: (a) our own limitations in understanding key natural and economic parameters and (b) genuine stochastic elements that govern the evolution of the systems under consideration. It can show up as unpredictable disturbances to the evolution of an ecosystem, either in the form of abrupt discrete occurrences (“catastrophic events”) or as an ongoing stream of small stochastic shocks which drive diffusion processes that need to be controlled.

Obviously, the diversity of uncertainty sources and types calls for a variety of methods to model and handle them as well as for various (often conflicting) policy measures to respond to their influence on the systems to be managed. Here we review various methods and approaches that have been considered in the literature for dealing with uncertainty in the context of natural resource management. We begin with a schematic (“canonical”) resource management model ([Sect. 47.2](#)) and proceed to show how the various types of uncertainty enter each of its elements ([Sect. 47.3](#)). In actual practice, resource managers may face more than a single type of uncertainty at the same time. We point out that the interaction between the various types can give rise to new complex effects.

In a more general setup, the management problem cannot be restricted to the resource sector but must be considered in a wider context, with various economy-wide variables both affecting and being affected by the environmental and natural

resource sectors. To account for such considerations, we describe a framework that integrates natural resources and aggregate economic growth and use it to discuss additional effects of uncertainty (Sect. 47.4). In Sect. 47.5, we direct attention to the concept of irreversibility characterizing many resource management situations. Irreversible outcomes are particularly relevant when coupled with uncertainty, because they can otherwise be anticipated and avoided when so desired. Finally, we discuss briefly the case of Knightian uncertainty (Sect. 47.6) under which the underlying structure of uncertainty (e.g., the specification of the underlying distribution) is incompletely known.

47.2 The Canonical Resource Management Model

In a typical resource management situation, an initial resource stock Q_0 is to be exploited over some planning horizon $t \in [0, T]$, t being the running time index and T is the end of the planning period which may or may not be predetermined. At any instant of time, the remaining stock $Q(t)$ is given, and the exploitation rate $q(t)$ generates the instantaneous benefit $u(Q(t), q(t), t)$ and changes $Q(t)$ according to

$$\dot{Q}(t) \equiv dQ(t)/dt = G(Q(t), q(t), t) \quad (47.1)$$

A simple example of a stock dynamic process is obtained from the specification $G(\cdot) = R(Q) - q$, where $R(\cdot)$ represents natural recharge (growth, replenishment). For nonrenewable resources, for example, minerals, R vanishes at all times and $G = -q$.

An exploitation policy $\{T, q(t), t \in [0, T]\}$ generates the payoff

$$\int_0^T u(Q(t), q(t), t) e^{-\rho t} dt + e^{-\rho T} v(Q(T)) \quad (47.2)$$

where ρ is the time rate of discount and $v(\cdot)$ is the post-planning value (the present value at time T of the benefit stream over the post-planning period $t > T$). The policy is feasible if it satisfies some given constraints on T and on $\{Q(t), q(t), t \in [0, T]\}$, for example, T is given or restricted to a certain range, the stock $Q(t)$ is positive or bounded in some range, and $q(t) \geq 0$ for all $t \in [0, T]$. We denote by Γ the set of all feasible policies.

The optimal policy is the feasible policy that maximizes Eq. (47.2) subject to Eq. (47.1) given $Q(0) = Q_0$. The value of Eq. (47.2) obtained under the optimal policy is denoted $V(Q_0; \Gamma)$ and is called the value function. For brevity, the argument Γ is often dropped, leaving the initial resource stock as the sole argument of the value function.

The formulation of the resource management problem in this way started with Hotelling (1931) who considered exhaustible (nonrenewable) resources and characterized optimal extraction policies in different market settings, using the calculus of variations to verify economic reasoning. The development of optimal control and dynamic programming methods opened the way for a wide range of extensions, including the incorporation of uncertainty of various kinds and forms.

In real-world situations, uncertainty is likely to be present in each of the components of the resource management problem: the planning horizon T , the instantaneous benefit $u(\cdot, \cdot, \cdot)$, the discount rate ρ , the post-planning value $v(\cdot, \cdot)$, the recharge process $R(\cdot, \cdot)$, the initial reserve Q_0 , as well as the specification of the feasibility constraints. In this chapter, we survey different approaches to deal with uncertainties often encountered in resource management problems.

Before delving into extensions involving uncertainty, it is expedient to summarize the salient properties of the optimal policy of the canonical management problem formulated above. Suppose that at some time t the resource owner is offered the opportunity to increase the remaining stock $Q(t)$ by a marginal unit. What is the maximal amount the owner will be willing to pay (at time t) to realize this opportunity? The answer, obviously, is the contribution of the added stock to the resource value at time t , that is, $V'(Q(t)) \equiv \partial V(Q)/\partial Q|_{Q=Q(t)}$. Let $\lambda(t)$ represent this opportunity cost at time t when the remaining stock is $Q(t)$. The variable $\lambda(\cdot)$ comes under various names, including costate, shadow price, scarcity or royalty rent, and in situ value. By definition, it embodies the economic implications of stock changes, such as increasing extraction costs as the resource dwindles and the price of scarcity when a nonrenewable resource is nearing depletion.

Exploitation at the rate $q(\cdot)$ bears two effects. First, it provides the instantaneous gratification $u(\cdot)$. Second, it changes the available stock via Eq. (47.1), hence the potential to enjoy future gratifications. The (current value) Hamiltonian,

$$H(Q, q, \lambda, t) \equiv u(Q, q, t) + \lambda G(Q, q, t)$$

balances these two effects such that the optimal exploitation rate maximizes it at each point of time. The economic interpretation of this “maximum principle” is readily seen under the specification $G(Q, q, t) = R(Q) - q$ and when the maximization admits an internal solution, in which case the optimal rate q satisfies $\partial u / \partial q = \lambda$: Along the optimal path, the marginal benefit from exploitation should equal the shadow price of the resource, that is, the marginal cost of exploitation.

Once the $\lambda(\cdot)$ process is given, the Hamiltonian maximization determines the optimal exploitation rate and, via Eq. (47.1), the ensuing stock process for the entire planning period $t \in [0, T]$. Solving the management problem, then, requires the determination of the shadow price process, for which optimal control and dynamic programming are two approaches.

In many cases, the optimal stock process $Q(\cdot)$ approaches a steady state (perhaps only asymptotically when $T = \infty$), where exploitation and natural recharge just balance each other out. This is the case, for example, in infinite horizon, autonomous problems (where the time argument enters explicitly only via discounting) involving a single stock. In such problems, it has been shown that the optimal stock process is monotonic, hence (when bounded) must eventually converge to a steady state. Deriving the steady state is relatively easy even for problems that do not admit analytic solutions for the full dynamic evolution. Comparing the steady states under different conditions (model specifications, parameter values) provides a simple way to study the effects of changes in the underlying conditions on the optimal policy.

The canonical resource management problem has been studied extensively, and the relevant literature is vast. For detailed treatments, we refer to Clark (1976) and Dasgupta and Heal (1979) who discussed resource management in a variety of situations, emphasizing renewable and nonrenewable resources, respectively.

47.3 Resource Management Under Uncertainty

As mentioned above, uncertainty abounds in resource management situations. It is important to distinguish at the outset between two types of uncertainty, depending on its origin. The first type is due to the participants' (resource owners, users, regulators, etc.) limited knowledge of certain parameters or functional relations characterizing the resource and the economic systems under consideration. The second type is due to genuine random elements often encountered when dealing with Mother Nature. We refer to the former type as *ignorance uncertainty* and to the latter as *exogenous uncertainty*. For example, the recharge or instantaneous benefit may undergo an abrupt shift when the stock process crosses some threshold, but the exact location of this threshold is a priori unknown. There is nothing inherently random in the threshold parameter, except that it is unknown to the resource manager; hence, the uncertainty is due to ignorance. If, however, the abrupt regime shift depends also on exogenous environmental factors such as weather variables affecting the outburst of a pollution-induced disease, then its occurrence is triggered by the confluence of environmental conditions which are genuinely stochastic, and the uncertainty regarding the abrupt shift is exogenous. How to handle a particular source of uncertainty depends to a large extent on its type.

We proceed now to discuss the incorporation of uncertainty, considering in turn each component of the above canonical resource management model.

47.3.1 Uncertain T

Some resource management problems do not admit a natural completion time, in which case the planning horizon becomes infinite ($T = \infty$). In other cases, extraction must cease at a finite date T , while the considerations related to later periods are summarized in the post-planning value $v(Q(T))$. For example, mine developers

may be permitted to extract the mineral only until some given date T when their concession expires. Moreover, the depletion of nonrenewable resources (or of renewable resources like fisheries that can be exploited to extinction) marks the end of the planning horizon, which depends on the extraction policy. In these cases, the planning horizon is either given exogenously or is a decision variable which can be determined for any extraction policy. In either case, its incorporation within the management problem involves no uncertainty and poses no particular difficulty.

In many situations, however, T is subject to uncertainty. A prominent example is that of an unknown initial stock – a situation studied initially by Kemp (1976). In such cases, T is a random variable whose realization marks the depletion of the resource, at which time management shifts to the post-planning period. A slight extension of the term “depletion” to include situations in which the resource can no longer be exploited or becomes obsolete allows to associate T with an uncertain date of nationalization (Long 1975) or with the uncertain arrival of a backstop substitute (Dasgupta and Heal 1974; Dasgupta and Stiglitz 1981). Cropper (1976) presented the problem in an environmental pollution context, identifying T with the random triggering of various environmental catastrophes.

While the uncertainty in the cake-eating problem of Kemp (1976) is solely due to ignorance, the uncertainty in political (nationalization) or economical (technological breakthrough) events often involves genuine stochastic elements and is therefore exogenous. The distinction between the two types of uncertainty plays out most pronouncedly via the specification of the hazard rate function, measuring the probability density of the event occurrence (the realization of T) in the next time instant. In all of these variants, the management problem seeks to maximize the *expected* value of the objective Eq. (47.2) with respect to the distribution of T , and the latter closely depends on the type of uncertainty.

47.3.1.1 Ignorance Uncertainty

A common ignorance-uncertainty situation involves a catastrophic event triggered by the stock falling below some unknown threshold. Examples, in addition to Kemp’s cake-eating problem, include seawater intrusion into coastal aquifers (Tsur and Zemel 1995) and global warming-induced catastrophes (Tsur and Zemel 1996; Nævdal 2006). The hazard rate in this case measures the probability of crossing the threshold during the next time instant. If the stock process does not decrease (e.g., extraction does not exceed the natural recharge) or if the stock process was in the past strictly lower than its current level, the hazard vanishes (it is certain that the threshold will not be crossed in the next time instant). In contrast, decreasing stock processes proceed under risk of occurrence. This feature complicates the formulation and solution of the management problem. The situation is greatly simplified if only monotonic stock processes are allowed. It turns out that in many cases of interest the *optimal* stock process is indeed monotonic.

The characterization of the optimal monotonic stock process proceeds along the following steps. Let \hat{Q}^c be the optimal steady state of the risk-free (canonical) problem. Consider an initial stock $Q_0 < \hat{Q}^c$. Since it is not optimal to decrease the stock further even without the risk of triggering a damaging event, it is obviously

not optimal to do so under the event risk. The optimal process under occurrence threat, then, coincides with the (increasing) risk-free process and approaches a steady state at \hat{Q}^c .

Suppose that $Q_0 > \hat{Q}^c$. Then, the optimal stock process cannot increase. For if it increases, the monotonicity property implies that it will never decrease, in which case the hazard vanishes at all times and the problem reduces to that of the risk-free problem. But without the occurrence risk, the optimal stock process converges to \hat{Q}^c – a contradiction. So when $Q_0 > \hat{Q}^c$, the optimal stock process is nonincreasing.

Let X denote the unknown threshold stock with the probability distribution $F(Q) \equiv \Pr\{X \leq Q\}$ and the corresponding density $f(Q) = F'(Q)$. For a decreasing stock process, the distribution

$$F_T(t) \equiv \Pr\{T \leq t\} = \Pr\{X \geq Q(t)\} = 1 - F(Q(t))$$

and the density

$$f_T(t) = F'_T(t) = -f(Q(t))\dot{Q}(t)$$

of the random occurrence time T determine the expected payoff (the expectation of Eq. (47.2) with respect to T). This expected payoff defines the objective of a deterministic management problem, denoted the “auxiliary” problem, which also admits a monotonic optimal stock process that converges to a steady state $\hat{Q}^{aux} > \hat{Q}^c$. It turns out that the resource management problem under uncertain threshold splits into two distinct deterministic subproblems, depending on the initial stock: For $Q_0 < \hat{Q}^c$, the optimal stock process is the same as the increasing *risk-free* process, and the occurrence risk can be ignored; for $Q_0 > \hat{Q}^{aux}$, the optimal process coincides with the decreasing *auxiliary* process, and the occurrence risk is relevant. If $Q_0 \in [\hat{Q}^c, \hat{Q}^{aux}]$, the uncertainty process enters a steady state instantly (at the initial state Q_0) because any other policy is ruled out by the above considerations. The *steady state interval* $[\hat{Q}^c, \hat{Q}^{aux}]$ is a peculiar feature, unique to optimal behavior under ignorance uncertainty.

Note the prudence implications of this characterization: Decreasing stock processes turn on the occurrence risk and hence approach a higher (and safer) steady state than that obtained without occurrence risk. Another interesting observation relates to the role of learning in this model. Decreasing stock processes provide new information regarding the threshold location as these processes proceed. This information, however, is already accounted for by the auxiliary objective, and the resource owners have no reason to update the original policy (designed at $t = 0$) as the information accumulates, unless the process is interrupted at some time by the catastrophic occurrence.

47.3.1.2 Exogenous Uncertainty

Under exogenous uncertainty, the event is triggered by genuinely random conditions, and the probability of occurrence within the next time instant is measured by the hazard rate (Long 1975; Cropper 1976; Heal 1984). The hazard rate in this case

depends neither on the history of the process nor on its trend (increasing or decreasing); hence, the splitting of the uncertainty problem into two distinct sub-problems (that gave rise to the equilibrium interval under ignorance uncertainty) does not occur. The hazard rate can, however, depend on the current resource stock and exploitation rate, which allows the owners to affect, even if not avoid completely, the risk of future occurrence by adjusting the extraction policy. This type of events has been assumed in a variety of resource models, including Deshmukh and Pliska (1985) who studied exploitation and exploration of nonrenewable resources, Reed and Heras (1992) in the context of biological resources vulnerable to a catastrophic collapse, Clarke and Reed (1994) and Tsur and Zemel (1998) in the context of pollution control, Cropper (1976) and Aronsson et al. (1998) who considered the risk of nuclear accidents, and Gjerde et al. (1999) and Bahn et al. (2008) in the context of climate policies under risk of environmental catastrophes.

Given the stock process $Q(\cdot)$, the stock-dependent hazard process $h(\cdot)$ is related to the probability distribution and density of the event occurrence time, $F(t) = \Pr\{T \leq t\}$ and $f(t) = F'(t)$, according to

$$h(Q(t))\Delta \equiv \Pr\{T \in (t, t + \Delta) | T > t\} = \frac{f(t)}{1 - F(t)}\Delta$$

Thus, $h(Q(t)) = -d \ln(1 - F(t))/dt$; hence,

$$F(t) = 1 - e^{-\int_0^t h(Q(s))ds} \quad \text{and} \quad f(t) = h(Q(t))[1 - F(t)]$$

The expectation (with respect to T) of the objective Eq. (47.2) becomes

$$\int_0^\infty [u(Q(t), q(t), t) + h(Q(t))v(Q(t))]e^{-\int_0^t [\rho + h(Q(\tau))]d\tau} dt \quad (47.3)$$

The optimal policy is the feasible policy that maximizes the objective Eq. (47.3) subject to Eq. (47.1) and $Q(0) = Q_0$. In this way, the uncertainty problem is recast as a standard deterministic infinite horizon problem. Its optimal policy is relevant only as long as the event has not occurred. Once the event occurs, the optimal policy switches to that of the post-event problem (represented by the post-event value v).

The event occurrence risk affects the resource management problem via the hazard rate, which enters the objective Eq. (47.3) both in the discount rate and in the instantaneous benefit ($u + hv$). The discount rate increases from ρ to $\rho + h$ with two conflicting effects. First, the increased impatience (due to the higher discount rate) promotes aggressive exploitation (less conservation). Second, the discount rate $\rho + h(Q)$ turns endogenous through its dependence on the stock. The possibility to control the discount rate via the extraction policy typically encourages conservation, and the trade-offs associated with the discounting effect are represented by the hazard rate of change $h'(Q)/h(Q)$.

The other effect of the occurrence threat on the management problem comes through the $h(Q)v(Q)$ term, which is added to the instantaneous benefit in the objective Eq. (47.3). When this term depends on the stock, the resource owners can control the expected damage of the event by adjusting the extraction policy. The overall uncertainty effect results from balancing these conflicting trends. In a particularly simple example, the post-event value $v(\cdot)$ vanishes identically at all Q levels. This is the case, for example, when the event occurrence renders the resource obsolete with no further consequences or when it is possible to renormalize the instantaneous benefit in such a way that the post-event value vanishes (see, e.g., Tsur and Zemel 2009; Karp and Tsur 2011). In this case, only the discounting effects remain. When the hazard is independent of the stock, only the impatience effect is active, and the ensuing optimal policy entails more aggressive exploitation than its risk-free counterpart: If the world may come to an end tomorrow and there is nothing we can do about it, we may as well exploit the resource today while we can. However, if the hazard is sensitive to the resource stock, such that more exploitation increases the occurrence probability, then the endogeneity of the discount rate encourages conservation. Which of these effects dominates depends on $h'(Q)/h(Q)$ (see discussion in de Zeeuw and Zemel 2012).

A slightly more general formulation describes the post-event value $v(\cdot)$ in terms of a penalty inflicted upon occurrence. Tsur and Zemel (1998) distinguish between single occurrence and “recurrent” events. The latter entails multiple penalties inflicted each (random) time the event occurs. For penalty functions that decrease with the stock, both types of events imply more conservative exploitation vis-à-vis the risk-free policy. A prominent example of recurrent events is the case of forest fires which affect forest rotation management (see Reed 1984).

Events that impact ecosystems often entail abrupt changes in the system dynamics. The post-event value in such cases is the outcome of the (risk-free) post-occurrence optimization problem proceeding under the new regime. When the change in dynamics implies a loss (e.g., via reduced natural replenishment of the resource), the extraction policy under uncertainty is more conservative than its risk-free counterpart (see Polasky et al. 2011 and references they cite). In fact, the discrete regime shift is in many cases a simplified description of the actual complex non-convex dynamical processes which give rise to fast transitions among locally stable basins of attraction and to hysteresis phenomena. However, when our interest is focused on the economic implications of the shift (rather than on the exact dynamics driving it), this simplification can yield interesting insights.

Catastrophic events of global nature, such as those induced by global warming, are often exogenous to local decision units (countries, regions). In such cases, the occurrence hazard is taken parametrically by the decision maker. The damage inflicted by the event, however, may change across locations, with particular grave outcomes to some specific nations. A possible response by local governments to this state of affairs is to consider adaptation activities in order to reduce or eliminate the damage that will be inflicted by the event, should the mitigation efforts (via reduced exploitation) fail to avoid its occurrence. The adaptation activities entail some given costs, while the benefit (of reduced damage) will be

enjoyed only following the (uncertain) occurrence date. The optimal adaptation policy should balance these costs and benefits (see de Zeeuw and Zemel 2012 and references therein). When the occurrence probability can be affected by mitigation policies, the two policy measures interact strongly and must be considered simultaneously to obtain optimal outcomes. Indeed, the mere presence of the adaptation option can modify the extraction policy even prior to the actual implementation of this option.

Our discussion has focused on unfavorable events such as environmental catastrophes. Favorable events, for example, technological breakthroughs, can be modeled in a similar way. Early studies of the uncertain arrival of a backstop substitute for nonrenewable resources with R&D efforts include Dasgupta et al. (1977), Kamien and Schwartz (1978) and Davison (1978). Bahn et al. (2008) considered such events in a renewable resource context of a climate policy that includes R&D efforts to develop clean energy technologies.

47.3.2 Stochastic Stock Dynamics

The dynamics of resource stocks is often driven by stochastic elements. Examples include biomass growth subject to random shocks, the replenishment of groundwater aquifers under uncertain precipitations, atmospheric pollution decay varying with changing weather conditions, and oil and mineral reserves subject to uncertain discoveries. The random shocks can come in the form of an ongoing stream of small fluctuations or as abrupt and substantial discrete occurrences. The latter show up, for example, when the resource evolution process undergoes a regime shift which entails the uncertain T scenario discussed above. Here we consider the continuous flow of small fluctuations giving rise to a diffusion (or random walk) process. As before, uncertainty regarding the stock evolution may be due to genuine random environmental shocks (Reed 1979; Pindyck 1984) or due to incomplete information. For example, the resource owners may be unable to measure the current stock precisely or to follow exactly the optimal extraction rule, leading to errors in predicting the next period's stock (Clark and Kirkwood 1986).

Reed (1974, 1979) considered a biomass stock (e.g., fish population) Q_t following the discrete-time natural growth rule

$$Q_{t+1} = Z_t R(Q_t)$$

where $R(\cdot)$ is the *expected* stock recruitment and Z_t are independently and identically distributed unit-mean random variables representing stochastic shocks affecting the population growth in each reproduction season. The resource stock is revealed following the realization of Z_t , yet the future evolution of the stock process cannot be predicted. In general, the concept of a steady state must be replaced by that of steady state distribution. However, if the realizations of the random shocks are observed before harvest decisions are made, the optimal policy maintains a constant escapement (postharvest biomass), that is, the optimal steady state

distribution of escapement degenerates to a constant (Reed 1979). When additional sources of uncertainty (e.g., errors in the measurement of current stocks) are added, the constant escapement rule no longer holds (see Sect. 47.3.6). A similar stochastic growth rule has been used by Weitzman (2002) to compare fishery regulation via landing fees with (the more common) harvest quota. He found that the former measure is more effective in this case. Observe that stochastic dynamics is not restricted to the population growth of some biological stock but might be relevant also to nonrenewable mineral stocks as a result of dedicated ongoing exploration efforts for new reserves with uncertain outcomes (Mangel 1985; Deshmukh and Pliska 1985).

Pindyck (1984) formulated the resource management problem under stochastic stock evolution in continuous time, employing Itô's stochastic calculus. The stock evolution follows a diffusion process which evolves according to the stochastic differential equation

$$dQ = [R(Q) - q]dt + \sigma(Q)dZ \quad (47.4)$$

where Z is a standard Wiener process and $\sigma^2(\cdot)$ is the corresponding variance. Specifying $\sigma(Q) = \sigma Q$, with σ a given constant, gives rise to a geometric Brownian motion and greatly facilitates the analysis. Taking again the expected cumulative net benefit as the objective for optimization, one can employ stochastic dynamic programming to derive the optimal extraction rule $q(Q)$ and the associated steady state distribution. The prudence implications for this type of uncertainty are again ambiguous and depend on the properties of the recharge and benefit functions (see Pindyck 1984 for examples in which the optimal exploitation rule $q(Q)$ increases, remains unchanged, or decreases as the variance parameter σ is increased).

Other examples of resource management under stochastic stock dynamics include Plourde and Yeung (1989), Knapp and Olson (1995), and Wirl (2006). The former considers pollution control when the accumulation process is stochastic due to the random absorption capacity of the ecosystem and finds that a user charge on inputs is preferable to the common “pollution standards” approach. This result is similar to that obtained by Weitzman (2002) in the discrete time setting. The second paper studies groundwater management with stochastic recharge due to uncertain precipitation, while the third studies climate policies under a stochastic global temperature process.

47.3.3 Discounting

Effects of discount rate variability are most pronounced when consequences of resource exploitation extend far into the distant future, such as in climate change or in nuclear waste disposal problems. In such cases, even slight changes in the discount rate entail exceedingly large differences in the weight assigned to the well-being of generations in the distant future and on optimal policies.

The discount rate changes with time preferences and technological shocks. Uncertain discounting due to future technological shocks has been analyzed in a number of works (see Gollier and Weitzman 2010 and references therein). Based on the discount rate distribution, an expression for the effective discount rate is derived and shown to decline gradually over time, approaching the lower end of the distribution in the long run. This feature can have large effects on optimal policies since it weighs the far future much more heavily than under the standard constant-rate discounting.

In light of the large variability observed in intragenerational time preferences, it is expected that the same holds for time preferences across generations. Thus, the time preferences of future generations are highly uncertain. These preferences depend on economic performance, technological progress, and availability of resources in the far future, and the treatment of the associated uncertainty requires integrating the canonical resource model of Sect. 47.2 within an economy-wide model. These issues are considered in Sect. 47.4.

47.3.4 Instantaneous Benefit

The flow of instantaneous benefit is also likely to be influenced by uncertain shocks, some of which are in the form of a stochastic diffusion process, while the others are substantial and abrupt. An example of the latter is a sudden drop of the demand for the resource as a result of a technological breakthrough (e.g., the effect of the development of fiber-optics communication on the demand for copper transmission lines). Such discrete shocks can be discussed in the context of uncertain time horizon T . A benefit diffusion process can be driven by a stochastic stock evolution (via the dependence of $u(\cdot)$ on the stock Q) as discussed in Sect. 47.3.2 or by benefit-specific fluctuations. An example of the latter is the stochastic demand for a nonrenewable resource introduced by Pindyck (1980).

Tsur and Graham-Tomasi (1991) studied renewable groundwater management when the demand for the resource fluctuates with rainfall. They distinguished between two information scenarios, depending on whether groundwater extraction decisions are made before or after the rainfall realization is observed. They also considered the reference case in which rainfall is stable at the mean. By comparing these three scenarios, they have been able to define the value of groundwater (the “buffer value”) due to its role in mitigating the fluctuations in water supply.

Conrad (1992) considered the control of stock pollutants when the pollution damage follows geometric Brownian motion, while Xepapadeas (1998) incorporated stochastic benefit shocks within a climate change model. The pollution stock process (atmospheric greenhouse gas concentration) is assumed to follow deterministic dynamics, but the damage it inflicts is modeled again as a diffusion process. The model considers a group of countries with deterministic private emissions and a stochastic public damage which depends on the global stock of pollution. The problem of coordinating emission abatement is analyzed via the optimal stopping methodology under cooperative and noncooperative modes of behavior on part of the participant countries.

47.3.5 Post-planning Value

The post-planning value determines the loss associated with occurrence hence the degree of effort that is optimally invested in avoiding the event or reducing its occurrence hazard. Uncertainty regarding this value is similar to that associated with the preplanning regime, such as uncertain post-planning stock dynamics or instantaneous benefit. For example, Goeschl and Perino (2009) study R&D efforts to develop a backstop substitute for a polluting resource. The exact nature of the substitute is subject to uncertainty, as it is not known in advance whether the backstop technology will also turn out eventually to be harmful to the environment (a “boomerang”) in which case yet another technology will need to be developed later on or it will solve the pollution problem for good. They show how the probability of either outcome affects the timing of adoption of the new technology.

Problems of long time horizons, such as global climate change, exacerbate the uncertainty regarding the post-planning value. Even if we knew precisely the temperature change a century ahead, it would be extremely hard to estimate the damage such a change would inflict on a future society which will surely differ greatly in its economic, technological, and demographic characteristics from what can be observed or predicted at the present time. Integrated assessment models, discussed in Sect. 47.4 below, deal with this kind of uncertainty in an ad hoc fashion.

47.3.6 Compound Uncertainties

The various uncertainty types presented above drive different responses in terms of the changes induced relative to the canonical certainty policy, with the *sign* of the change depending on the particular type under consideration. It is often of interest to study how the *magnitude* of these changes depends on uncertainty, when the latter is measured, for instance, by the variance of a related key parameter (e.g., the parameter σ^2 of Eq. (47.4)). Typically, each source of uncertainty drives the policy along a well-defined trend, and the effect responds monotonically to changing uncertainty. However, many resource management problems are subject to the combination of more than one type of uncertainty. When two (or more) types of uncertainty are combined, the policy response becomes more involved than in the case of a single type because the interaction between the types can give rise to new phenomena. Aiming to account for such situations, Clark and Kirkwood (1986) combined Reed’s (1979) discrete stochastic fish stock dynamics with measurement errors on the stock size at the beginning of each harvesting period, while Sethi et al. (2005) added a third component, namely, the inaccurate implementation of the harvest policy in each period. They showed that Reed’s (1979) constant escapement rule is no longer optimal when harvest decisions are made before realizations of the random shocks are observed, in which case the optimal policy may not admit analytic solution and the planner must resort to numerical methods.

The effect of the interactions among different types of uncertainty is evident in the work of Saphores (2003) who considered stochastic stock dynamics under the threat of extinction if the biomass hits a barrier and found a non-monotonic response to increasing the stochastic variance: The increase in variance implies more precaution when the variance is small but calls for more aggressive harvesting when the variance is large enough. More recently, Brozović and Schlenker (2011) obtained a similar outcome when the stochastic stock dynamics is combined with the risk of an abrupt shift in ecosystem dynamics. These models allow the planner to take actions at discrete points of time, and the non-monotonic behavior is attributed to changes implied by increasing the variance on the trade-off between reducing the shift probability vs. the cost of precautionary behavior.

Leizarowitz and Tsur (2012) studied optimal management of a stochastically replenished (or growing) resource under threat of a catastrophic event such as eutrophication (of shallow lakes), species extinction, or ecosystem collapse. They considered discrete time and discrete state and action spaces. The catastrophic threat renders the single-period discount factor policy dependent, and as a result the compound discount factor becomes history dependent. The authors investigated whether an optimal Markovian-deterministic stationary policy exists for this problem. They answered this question in the affirmative and verified that the optimal state process converges to a steady state distribution. They identified cases under which the steady state distribution implies that the event will eventually occur with probability one and contrasted them with cases under which the catastrophic event will never occur.

Employing a continuous-time formulation, Yin and Newman (1996) combined a stochastic output price process (as in Conrad 1992) with the catastrophic forest fires of Reed (1984) and found that the risk of fire entails different responses depending on whether the fire is a single event that prevents further exploitation or investments and fires can reoccur. In a similar framework, Balikcioglu, Fackler, and Pindyck (2011 and references therein) combined the stochastic pollution stock dynamics (analogous to Eq. (47.4)) with stochastic uncertainty regarding the damage inflicted by this stock (as in Xepapadeas 1998). The optimal response is analyzed again via stopping theory, and the complexity introduced by the dual source of uncertainty necessitates the development of a sophisticated numerical method of solution.

Zemel (2012) provides an analytic, continuous-time confirmation of the non-monotonic response by incorporating the uncertain regime shifts of de Zeeuw and Zemel (2012) into the stochastic stock model Eq. (47.4). It is verified that the simultaneous action of both types of uncertainty is indeed required to obtain this behavior. When one or the other sources of uncertainty are switched off, the other acts to promote conservation (as expected). However, when the two sources interact, increasing the stochastic variance enhances the hazard effect when the variance is small but works in the opposite direction when the variance is large. In a world of multiple sources of uncertainty, it is therefore likely that non-monotonic response is more common than the simple, single-uncertainty-type models would suggest.

Obviously, combining several uncertainty sources greatly complicates the management problem, and one usually has to resort to numerical methods to derive the optimal policy. This is the approach adopted by the integrated assessment models discussed in Sect. 47.4 below.

47.4 Integrating Natural Resources and Aggregate Growth Models

Some uncertain elements affect resource exploitation indirectly via their influence on economy-wide variables. Examples include the intra- and intergenerational variability of time preferences and technological shocks. Accounting for these uncertain elements requires incorporating the canonical resource model of Sect. 47.2 within an economy-wide (growth) framework. The approach taken in this section is in line with the views of ecological economists who have pointed out that problems of economic growth cannot be decoupled from the constraints imposed by the embedding environmental system. We briefly outline an integrated model of this kind and use it to discuss additional effects of uncertainty.

47.4.1 An Integrated Model

An important (though not the only) role of natural resources is to serve as sources of production inputs. Accordingly, suppose the extracted resource q is used as an input of production alongside capital K and human capital augmented labor AL (A is an index of human capital and L represents the labor force) to produce the output Y according to the technology $Y = F(K, q, AL)$. The wealth of an economy is measured by its stocks of natural capital Q , producible capital K , and human capital A . The former changes according to Eq. (47.1) and K changes according to

$$\dot{K} = F(K, q, AL) - C - z(Q, q) - \delta K \quad (47.5)$$

where C is aggregate consumption, δ is a depreciation parameter, and $z(\cdot)$ is the extraction cost. (In the canonical model of Sect. 47.2, $z(Q, q)$ is embedded in the instantaneous benefit $u(Q, q, t)$, which is here replaced by the consumption utility.) The evolution of human capital may be driven by exogenous labor-augmenting technical change processes or by endogenous policies. Equation (47.5), then, can be viewed as a variant of a Solow-type growth model.

Per capita consumption, $c = C/L$, generates the per capita instantaneous utility $u(c)$, and welfare is measured by the present value of the utility stream

$$\int_0^T Lu(c)e^{-\rho t} dt + e^{-\rho T} v(Q(T)) \quad (47.6)$$

where ρ is the utility discount rate which discounts future consumption solely due to the passage of time and should be distinguished from the interest rate r (the price of capital). The resource allocation problem requires to find the feasible consumption-exploitation-investment policy that maximizes the welfare Eq. (47.6) subject to the dynamic evolution of the capital stocks, given the endowment Q_0 , K_0 , and A_0 . More general variants of this model allow for multiple resources and for an explicit dependence of the utility also on some of the stocks (e.g., a clean environment or the preservation of species; see Heal and Kriström 2002 and references therein)

In equilibrium the optimal policy follows (under some conditions) Ramsey's formula $r = \rho + \eta g$, where η is the elasticity of marginal utility and g is the rate of growth of per capita consumption. This condition varies with intergenerational variations in preferences (ρ and η) and in the growth rate (g). The "correct" rate to be used is controversial (see Stern 2008; Nordhaus 2008, and references therein), and the controversy is exacerbated by the uncertain future evolution of these variables.

47.4.2 Uncertainty in the Integrated Model

The integrated model allows us to address a wider range of uncertainties as well as to study feedback effects between natural resources and the wider economy. For example, Tsur and Zemel (2009) looked at the effect of economic growth on climate policy regarding greenhouse gas (GHG) emission under threat of a catastrophic climate change whose occurrence probability depends on atmospheric GHG concentration. They found that economic growth motivates more vigorous mitigation of GHG emission such that in the long run anthropogenic GHG emission (beyond the natural rate) should be banned altogether. The reason is rather straightforward: As the economy grows richer, it stands to lose more in case the catastrophe strikes, while at the same time it can more easily afford to relinquish the resources needed to use and develop clean substitutes. What is less obvious is that, due to the global public bad nature of the threat induced by atmospheric GHG concentration, the market outcome gives rise to the opposite allocation, namely, maximal (in economic terms) use of polluting fossil fuels. Such an interaction between an economy-wide phenomenon, in the form of economic growth induced by technical change, and resource exploitation affecting the probability of triggering a damaging event can be addressed only within an integrated framework.

As integrated models (particularly those aiming at describing faithfully the real world) tend to be analytically intractable, they call for the use of numerical analysis. Examples are the so-called integrated assessment models that link together climate and aggregate growth models (see Stern 2008; Nordhaus 2008, and references therein). Uncertainty in these models is often treated by considering a distribution for each of the unknown parameters and deriving the results for a large number of "scenarios", each corresponding to a particular parameter specification. The results are then reported in terms of the most likely values as well as of some measure of their spread.

47.5 Irreversibility and Uncertainty

A ubiquitous feature of environmental management problems is the irreversibility characterizing many natural processes. This feature can come in the form of the abrupt catastrophic occurrences discussed above (examples of which are the reversal of the flow of the Gulf Stream due to global temperature rise, species extinction due to overharvesting or habitat destruction, the collapse of groundwater aquifers due to seawater intrusion, or the eutrophication of lakes as a result of the use of fertilizers along their shores). Otherwise, some of our actions (polluting emissions, forest clearing, or the extraction of exhaustible resources) cannot be undone (or can be corrected very slowly) when an unfavorable outcome is realized. These irreversible regime shifts are manifestations of non-convexities in the dynamic equations that drive the underlying natural processes. This feature implies fast transitions among competing stable equilibria and hysteresis phenomena. As stated above, the simplified description of these phenomena as irreversible transitions provides a useful approximation to derive the management policies.

The presence of irreversibility really matters only under uncertainty, because otherwise undesirable outcomes can be anticipated in advance and avoided. Heal and Kriström (2002), Pindyck (2007), and the references they cite discuss in detail the effect of irreversibility on management policies under uncertainty. Presenting the problem in terms of the theory of real options, they identify two diametrical effects. If the damage associated with occurrence will turn out in the future to be very large, then exercising the option of aggressive extraction today entails a significant social loss. This effect pushes the cost-benefit balance towards more conservation. However, abatement activities often involve sunk costs (e.g., the purchase of abatement equipment that can be used only for that purpose) which give rise to the opposite effect. If it eventually turns out that the occurrence hazard or the associated damage has been overestimated, the abatement investment cannot be undone, and failing to exercise the option to wait and learn more about the hovering threat might turn out costly.

The irreversibility-induced trade-offs are particularly pronounced in optimal stopping problems (e.g., Balikcioglu et al. 2011 and the references they cite) where the problem is to determine the optimal time to enact an irreversible change in policy (e.g., reduce emissions) at a sunk cost when the pollution and damage processes follow stochastic dynamics. This regime shift problem is reminiscent of the uncertain regime shift time T discussed in Sect. 47.3.1. Here, however, the time of shift is the decision variable rather than an exogenous parameter subject to uncertainty. Optimal stopping has also been used to study the optimal time to invest in R&D efforts aimed at developing a substitute for a nonrenewable resource (Hung and Quyen 1993) or for a polluting technology (Goeschl and Perino 2009).

Wirl (2006) considered the consequences of two types of irreversibility on optimal CO₂ emission policies when the temperature follows a diffusion process. First, emissions are irreversible in the sense that active collection of the polluting gases out of the atmosphere is not allowed. Moreover, stopping is irreversible so that once the decision to stop emissions is taken, it cannot be reversed. He found that these effects work against conservation and that irreversible stopping is never optimal.

47.6 Knightian Uncertainty

The literature cited so far treats uncertainty by converting random variables into expectations based on well-specified distribution functions. Often, however, the distribution functions themselves are only partially known – a situation referred to as Knightian (or structural) uncertainty. For example, as perceived at present, future growth rates may be random with unknown mean and/or standard deviation. When realizations of an informative random variable are progressively observed, the underlying distribution can be deduced via Bayesian updating with progressive levels of accuracy.

However, if the downside of possible outcomes (e.g., the consequences of a climate change induced catastrophe) is not bounded, the expected present value may be unbounded as well for any incomplete information (finite number of observations) underlying the Bayesian updated (posterior) probabilities. This situation was illustrated by Weitzman (2009) in a two-period model in which growth is random (due to a random climate parameter) with a distribution that is known only up to a scale parameter. The analysis points to the potential limitations of combining expected utility theory and Bayesian updating in analyzing decisions under uncertainty in general and for resource management in particular. Alternative approaches, involving the precautionary principle and ambiguity-averse learners, have recently been considered for resource management problems (see Vardas and Xepapadeas 2010 and references therein).

47.7 Conclusions

The proper response to uncertainty has become a prevailing consideration in the resource management literature, and the survey in this chapter attempts to expose the diversity of approaches developed for this purpose. A necessary step in dealing with uncertainty is the recognition that uncertainty is present in nearly every aspect of a resource management problem and that different types of uncertainty call for policy responses that may differ substantially and in some cases even diametrically. For example, some types of uncertainty encourage more conservation and cautious exploitation, while other types induce the opposite response – of a more vigorous exploitation (relative to the comparable situation managed under certainty).

Although our aim was to cover the wide range of stochastic aspects relevant for environmental and natural resources management, it is recognized that a comprehensive treatment is not feasible within the limits of a single chapter and some important aspects had to be left out. For example, environmental resources are often shared by several agents, and their management is subject to strategic interactions among competing stake holders. These interactions are usually studied via the theory of dynamic games and involve again uncertainty of various types, including that due to asymmetric information among players (see Dockner et al. 2000 and the literature cited therein). The treatment of this important and complex topic is beyond the scope of this chapter.

References

- Aronsson T, Backlund K, Löfgren KG (1998) Nuclear power, externalities and non-standard pigouvian taxes: a dynamic analysis under uncertainty. *Environ Resour Econ* 11:177–195
- Bahn O, Haurie A, Malhamé R (2008) A stochastic control model for optimal timing of climate policies. *Automatica* 44:1545–1558
- Balikcioglu M, Fackler PL, Pindyck RS (2011) Solving optimal timing problems in environmental economics. *Resour Energy Econ* 33:761–768
- Brozović N, Schlenker W (2011) Optimal management of an ecosystem with an unknown threshold. *Ecol Econ* 70:627–640
- Clark CW (1976) Mathematical bioeconomics: the optimal management of renewable resources. Wiley, New York
- Clark CW, Kirkwood GP (1986) On uncertain renewable resource stocks: optimal harvest policies and the value of stock surveys. *J Environ Manag* 13:235–244
- Clarke HR, Reed WJ (1994) Consumption/pollution tradeoffs in an environment vulnerable to pollution-related catastrophic collapse. *J Econ Dyn Control* 18:991–1010
- Conrad JM (1992) Stopping rules and the control of stock pollutants. *Natural Resour Model* 6:315–327
- Cropper ML (1976) Regulating activities with catastrophic environmental effects. *J Environ Econ Manag* 3:1–15
- Dasgupta P, Heal G (1974) The optimal depletion of exhaustible resources. *Rev Econ Stud* 41:3–28
- Dasgupta P, Heal GM (1979) Economic theory and exhaustible resources. Cambridge University Press, Cambridge
- Dasgupta P, Stiglitz J (1981) Resource depletion under technological uncertainty. *Econometrica* 49:85–104
- Dasgupta P, Heal G, Majumdar M (1977) Resource depletion and research and development. In: Intriligator MD (ed) *Frontiers of quantitative economics*, vol III B. North-Holland, Amsterdam
- Davison R (1978) Optimal depletion of an exhaustible resource with research and development towards an alternative technology. *Rev Econ Stud* 45:355–367
- de Zeeuw A, Zemel A (2012) Regime shifts and uncertainty in pollution control. *J Econ Dyn Control* 36:939–950
- Deshmukh SD, Pliska SR (1985) A martingale characterization of the price of a nonrenewable resource with decisions involving uncertainty. *J Econ Theory* 35:322–342
- Dockner EJ, Jorgensen S, Long NV, Sorger G (2000) Differential games in economics and management science. Cambridge University Press, Cambridge
- Gjerde J, Grepperud S, Kverndokk S (1999) Optimal climate policy under the possibility of a catastrophe. *Resour Energy Econ* 21:289–317
- Goeschl T, Perino G (2009) On backstops and boomerangs: environmental R&D under technological uncertainty. *Energy Econ* 31:800–809
- Gollier C, Weitzman ML (2010) How should the distant future be discounted when discount rates are uncertain? *Econ Lett* 107:350–353
- Heal G (1984) Interactions between economy and climate: a framework for policy design under uncertainty. *Adv Appl Microecon* 3:151–168
- Heal G, Kriström B (2002) Uncertainty and climate change. *Environ Resour Econ* 22:3–39
- Hotelling H (1931) The economics of exhaustible resources. *J Polit Econ* 39:137–175
- Hung NM, Quyen NV (1993) On R&D timing under uncertainty: the case of exhaustible resource substitution. *J Econ Dyn Control* 17:971–991
- Kamien MI, Schwartz NL (1978) Optimal exhaustible resource depletion with endogenous technical change. *Rev Econ Stud* 45:179–196
- Karp L, Tsur Y (2011) Time perspective and climate change policy. *J Environ Econ Manag* 62:1–14
- Kemp MC (1976) How to eat a cake of unknown size. In: Kemp MC (ed) *Three topics in the theory of international trade*. North-Holland, Amsterdam

- Knapp K, Olson L (1995) The economics of conjunctive groundwater management with stochastic surface supplies. *J Environ Econ Manag* 28:340–356
- Leizarowitz A, Tsur Y (2012) Renewable resource management with stochastic recharge and environmental threats. *J Econ Dyn Control* 36:736–753
- Long NV (1975) Resource extraction under the uncertainty about possible nationalization. *J Econ Theory* 10:42–53
- Mangel M (1985) Decision and control in uncertain resource systems. Academic Press, Orlando
- Nævdal E (2006) Dynamic optimization in the presence of threshold effects when the location of the threshold is uncertain – with an application to a possible disintegration of the western antarctic ice sheet. *J Econ Dyn Control* 30:1131–1158
- Nordhaus WD (2008) A question of balance: weighing the options on global warming policies. Yale University Press, New Haven
- Pindyck RS (1980) Uncertainty and exhaustible resource markets. *J Polit Econ* 88:1203–1225
- Pindyck RS (1984) Uncertainty in the theory of renewable resource markets. *Rev Econ Stud* 51:289–303
- Pindyck RS (2007) Uncertainty in environmental economics. *Rev Environ Econ Policy* 1:45–65
- Plourde C, Yeung D (1989) A model of industrial pollution in a stochastic environment. *J Environ Econ Manag* 16:97–105
- Polasky S, de Zeeuw A, Wagener F (2011) Optimal management with potential regime shifts. *J Environ Econ Manag* 62:229–240
- Reed WJ (1974) A stochastic model for the economic management of a renewable animal resource. *Math Biosci* 22:313–337
- Reed WJ (1979) Optimal escapement levels in stochastic and deterministic harvesting models. *J Environ Econ Manag* 6:350–363
- Reed WJ (1984) The effect of the risk of fire on the optimal rotation of a forest. *J Environ Econ Manag* 11:180–190
- Reed WJ, Heras HE (1992) The conservation and exploitation of vulnerable resources. *Bull Math Biol* 54:185–207
- Saphores JD (2003) Harvesting a renewable resource under uncertainty. *J Econ Dyn Control* 28:509–529
- Sethi G, Costello C, Fisher A, Hanemann M, Karp L (2005) Fishery management under multiple uncertainty. *J Environ Econ Manag* 50:300–318
- Stern N (2008) The economics of climate change. *Am Econ Rev* 98:1–37
- Tsur Y, Graham-Tomasi T (1991) The buffer value of groundwater with stochastic surface water supplies. *J Environ Econ Manag* 21:201–224
- Tsur Y, Zemel A (1995) Uncertainty and irreversibility in groundwater resource management. *J Environ Econ Manag* 29:149–161
- Tsur Y, Zemel A (1996) Accounting for global warming risks: resource management under event uncertainty. *J Econ Dyn Control* 20:1289–1305
- Tsur Y, Zemel A (1998) Pollution control in an uncertain environment. *J Econ Dyn Control* 22:967–975
- Tsur Y, Zemel A (2009) Endogenous discounting and climate policy. *Environ Resour Econ* 44:507–520
- Vardas G, Xepapadeas A (2010) Model uncertainty, ambiguity and the precautionary principle: Implications for biodiversity management. *Environ Resour Econ* 45:379–404
- Weitzman ML (2002) Landing fees vs harvest quotas with uncertain fish stocks. *J Environ Econ Manag* 43:325–338
- Weitzman ML (2009) On modeling and interpreting the economics of catastrophic climate change. *Rev Econ Stat* 91:1–19
- Wirl F (2006) Consequences of irreversibilities on optimal intertemporal CO₂ emission policies under uncertainty. *Resour Energy Econ* 28:105–123

- Xepapadeas A (1998) Policy adoption rules and global warming. *Environ Resour Econ* 11:635–646
- Yin R, Newman DH (1996) The effect of catastrophic risk on forest investment decisions. *J Environ Econ Manag* 31:186–197
- Zemel A (2012) Precaution under mixed uncertainty: implications for environmental management. *Resour Energy Econ* 34:188–197

Hassan Bencheikroun and Ngo Van Long

Contents

48.1	Introduction	951
48.2	Static Games	952
48.2.1	The Emissions Game	953
48.2.2	Sustaining Cooperation in a Noncooperative World	954
48.3	Dynamic Games: Some Concepts	959
48.4	Transboundary Stock Pollutants	963
48.4.1	A Benchmark Model	963
48.4.2	Centralized Versus Regional Control of Pollution	967
48.5	Provision of Clean Air and Interregional Mobility of Capital	968
48.6	Conclusions	969
	References	970

Abstract

We cover applications of game theory in environmental and resource economics with a particular emphasis on noncooperative transboundary pollution and resource games. Both flow and stock pollutants are considered. Equilibrium concepts in static and dynamic games are reviewed. We present an application of game theoretical tools related to the formation and sustainability of cooperation in transboundary pollution games. We discuss the analytical tools relevant for the case of a stock pollutant and offer an application related to the optimal institutional arrangement to regulate a pollutant when several jurisdictions are involved.

H. Bencheikroun (✉) • N.V. Long

Department of Economics and CIREQ, McGill University, Montréal, QC, Canada
e-mail: Hassan.bencheikroun@mcgill.ca; ngo.long@mcgill.ca

48.1 Introduction

Static and dynamic games have offered important tools to study many strategic interactions in natural resource and environmental economics as well as regional science and management science. The main difference between static games and dynamic games is that the latter deal with situations where economic agents operate in an environment that changes over time and agents can influence the evolution of the environment.

In analyzing any problem of strategic interactions, it is usually better to begin with the simplest model. This often means that one should, as a first step, abstract from dynamic considerations. Static game theory is sufficiently rich to shed lights on many scenarios of social and economic interactions. On the other hand, many problems in economics are temporal problems by nature, and eventually the temporal dimension must be taken into account. For this reason, dynamic game models are often encountered in scientific journals in fields such as resource and environmental economics and regional and urban economics.

Some warnings are in order. The environmental economics literature with a special interest in strategic behavior between regions is large. Since this chapter seeks to be self-contained, and given the space limitation, the material presented should be seen as a sample of the application of game theoretic tools to important classes of regional environmental and resource economics problems in a multi-region context. In particular we shall omit applications of cooperative game theory and only present a selection of noncooperative game theoretic models. For recent surveys of applications of game theory in environmental economics, we refer the reader to Jorgensen et al. (2010) and Long (2011).

In Sect. 48.2, we consider within a static model the issue of environmental agreements. In Sect. 48.3, we turn to dynamic games with simultaneous moves and briefly explain various equilibrium concepts in dynamic games. In Sects. 48.4 and 48.5, we provide illustrations of these concepts applied to problems of natural resources and environmental economics with a multi-region setting.

48.2 Static Games

Static game theory is better suited for transboundary pollution problems involving emission and where a few players interact strategically. We present an application of game theoretical tools related to the formation and sustainability of cooperation in pollution games. We review emission games and abatement games and the comprehensive analytical treatment by Rubio and Ulph (2006) of the canonical model of international environmental agreements initiated by Barrett (1994).

We begin with a noncooperative game of emissions. Then we turn to the question of how cooperation can be achieved and an analysis of stable coalitions.

Note that while the models are presented in the case interaction among countries, they apply also to the case of a single country made of several regions with autonomous regulatory powers over pollution and resource use, as is the case in

many countries. The main feature of these problems is the absence of a supranational authority or the lack of constitutional power of a central authority such as a federal government.

48.2.1 The Emissions Game

Consider a world consisting of N countries $i = 1, \dots, N$. A strategy for country i is a nonpositive level of emissions $q_i \geq 0$. Country i derives a net benefit flow

$$\pi_i(q_i, Q_i) \equiv aq_i - \frac{b}{2}q_i^2 - \frac{1}{2}(q_i + Q_{-i})^2$$

where $Q_{-i} \equiv \sum_{k \neq i} q_k$ and a and b are two positive parameters. The term $aq_i - \frac{b}{2}q_i^2$ measures the gross benefit from consumption, and the term $\frac{1}{2}(q_i + Q_{-i})^2$ measures the environmental damages each country suffers from the total emissions Q . Note that since the marginal damage from emissions is normalized to one, a large value of b represents a large marginal benefit or a small marginal damage cost.

Assuming countries choose their actions simultaneously, the unique Nash equilibrium strategy is

$$q_i = q^* = \frac{a}{b + N}$$

and the equilibrium payoff is

$$\pi^* = \frac{1}{2} \frac{(-N^2 + 2N + b)}{(N + b)^2} a^2$$

Let q^c denote the level of emissions that maximizes world welfare. Then,

$$q^c = \frac{a}{b + N^2} < q^* = \frac{a}{b + N}$$

Clearly, welfare under cooperation is higher than Nash equilibrium level,

$$\pi^c = \frac{1}{2} \frac{a^2}{N^2 + b} > \pi^* = \frac{1}{2} \frac{(-N^2 + 2N + b)}{(N + b)^2} a^2$$

$$\pi^c - \pi^* = \frac{1}{2} \frac{N^2(N - 1)^2}{(N^2 + b)(N + b)^2} a^2$$

Note that the gains from cooperation is decreasing in b . The gains from cooperation are most substantial when $b \rightarrow 0$.

The above game, analyzed in Rubio and Ulph (2006), is a game of pollution emissions, which can be compared with the abatement game of Barrett (1994). (There exists a correspondence between the emissions game and an abatement game (see Appendix 1 in Rubio and Ulph (2006)).)

48.2.2 Sustaining Cooperation in a Noncooperative World

We have shown that the noncooperative outcome is inefficient. Let us consider a possible improvement by some form of cooperation. Suppose a subgroup of the players considers coordination of their strategies to improve on their noncooperative equilibrium payoff. We define an international environmental agreement (IEA) as cooperation among M countries, where $M \geq 2$. Assuming the nonexistence of a supranational authority, we require any agreement to improve on the Nash equilibrium outcome to be self-sustaining.

We formulate an IEA game as a metagame where an emissions game (Or an abatement game) is preceded by an initial stage where countries decide whether to join a coalition or not. An IEA consisting of $M \leq N$ members chooses a vector of M strategies, one for each coalition member, to maximize the sum of their payoffs. When $M = N$, the coalition is called the grand coalition.

Several criteria of stability have been proposed in the theory of coalition formation. The predominant stability criterion in the IEA literature uses the concepts of internal and external stability. This criterion is based on the assumption that when a country considers the gain from defection, it supposes that all countries in the IEA would continue to cooperate and maximize their joint welfare. (An alternative stability criterion is that of farsighted stability (for more details and references, see Benchekroun and Long (2011) Sect. 48.4.2).)

Let $\pi^s(M)$ and $\pi^{ns}(M)$ denote the equilibrium payoffs of the representative signatory and non-signatory countries. We say that a given IEA with M members is internally stable if no signatory gains by leaving the IEA, i.e., $\pi^s(M) \geq \pi^{ns}(M - 1)$. Similarly, external stability means that a non-signatory does not gain by joining the IEA, i.e., $\pi^{ns}(M) \geq \pi^s(M + 1)$. An IEA is stable if and only if it is both internally and externally stable.

Once an IEA has been formed, in stage 2 game, one may consider two scenarios: (1) IEA members and non-signatories choose their actions simultaneously or (2) IEA members are the first movers, announcing and committing to their emissions policies before non-signatories can act. Most papers in the literature prefer the second scenario, i.e., the IEA members play a leadership role in the emissions game. We report below the analysis of the second scenario, following Rubio and Ulph (2006).

48.2.2.1 Stage 2: The Emissions Game Revisited

Using backward induction, let us first determine the reaction function non-signatories. Suppose the first M countries are signatories. A non-signatory country $k > M$ seeks to

$$\max_{q_k \geq 0} \left(aq_k - \frac{b}{2}q_k^2 - \frac{1}{2}(q_k + Q_{-k})^2 \right)$$

Its reaction function is

$$q_k = \max \left\{ \frac{a - Q_{-k}}{b + 1}, 0 \right\} \quad (48.1)$$

Knowing the reaction function of the non-signatories, the collection of signatories chooses their emissions to maximize the sum of their payoffs,

$$\max_{q_1, \dots, q_M \geq 0} \sum_{i=1}^M \left(aq_i - \frac{b}{2}q_i^2 - \frac{1}{2}(q_i + Q_{-i})^2 \right)$$

subject to the reaction function of the non-signatories. Under the symmetry assumption among coalition members, the maximization problem becomes

$$\max_{q_s \geq 0} M \left(aq_s - \frac{b}{2}q_s^2 - \frac{1}{2}(Mq_s + (N - M)q_{ns})^2 \right)$$

subject to

$$q_{ns} = \max \left\{ \frac{a - Mq_s}{b + N - M}, 0 \right\}$$

where q_s and q_{ns} denote respectively the emissions of a signatory and a non-signatory country.

Following Rubio and Ulph (2006), consider three possibilities, depending on interior or corner solutions. For this purpose, define

$$g(b, M) \equiv b^2 - (N - M)(M - 2)b + (N - M)^2$$

and

$$h(b, M) \equiv b^2 + (N + M^2 - 2M)b - (N - M)M$$

The three possible cases are as follows.

- (a) Interior solutions for all countries. This occurs if and only if $g(b, M) > 0$ and $h(b, M) > 0$. The interior solutions are given by

$$q_s = \frac{ag(b, M)}{b\omega}$$

and

$$q_{ns} = \frac{ah(b, M)}{b\omega}$$

where

$$\omega \equiv \left((b + N - M)^2 + bM^2 \right)$$

The equilibrium payoffs are given by

$$\pi^s(M) = \frac{a^2}{2b} \left(1 - \frac{bN^2}{\omega^2} \right)$$

and

$$\pi^{ns}(M) = \frac{a^2}{2b} \left(1 - \frac{(b+1)N^2 (b+N-M)^2}{\omega^2} \right)$$

(b) Corner solution for signatories. This occurs when $g(b, M) \leq 0$. Then $q_s = 0$ and

$$q_{ns} = \frac{a}{b + N - M}$$

with equilibrium payoffs

$$\pi^s(M) = - \frac{a^2 (N - M)^2}{2 (b + N - M)^2}$$

and

$$\pi^{ns}(M) = \frac{a^2 (b - (N - M)(N - M - 2))^2}{2 (b + N - M)^2}$$

(c) Corner solution for non-signatories. This occurs if $h(b, M) \leq 0$. Then $q_{ns} = 0$ and

$$q_s = \frac{a}{M}$$

with equilibrium payoffs

$$\pi^s(M) = - \frac{a^2 (b + M(M - 2))}{2M^2}$$

and

$$\pi^{ns}(M) = - \frac{a^2}{2}$$

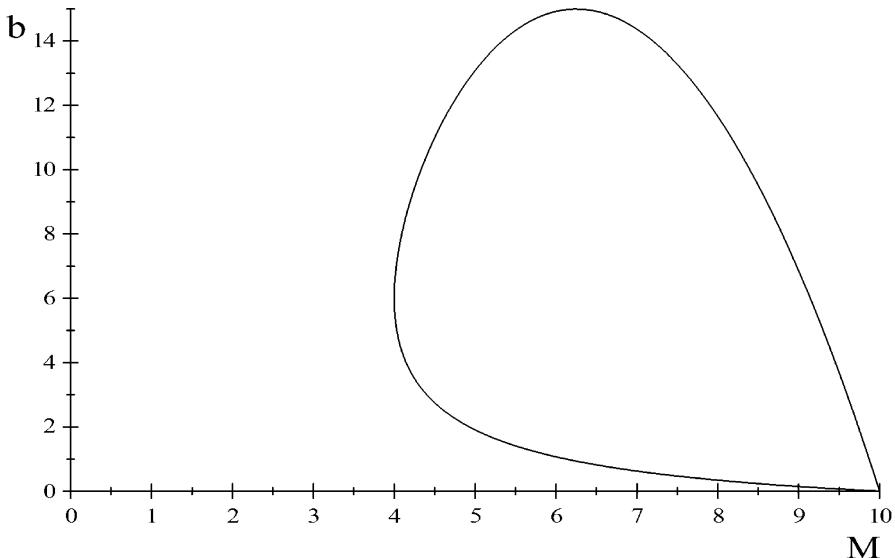


Fig. 48.1 Contour plot of $g(b, M)$ when $N=10$

Since functions g and h cannot be both negative, the configuration $q_s = q_{ns} = 0$ is not possible in a Stackelberg equilibrium. This is because the marginal benefit at zero pollution level is equal to $a > 0$, whereas the marginal damage of pollution at zero is nil. Finally, note that when $M = 1$ or $M = N$, the solution is interior.

Which solution occurs depends on b , N , and M . The creation of a coalition or a change in the size of a coalition can result in a change from an interior to a corner solution or vice versa. Therefore, before tackling the issue of coalition formation, it is important to clarify how the signs of the functions g and h depend on the model parameters. Rubio and Ulph fix N and study the sign of g and h as a function of b and M . The analysis can be summed up in the figures below. Figure 48.1 depicts the level curve $g(b, M) = 0$. The interior of convex region depicted represents all (b, M) such that non-signatories choose zero emissions. Figure 48.2 depicts the level curve $h(b, M) = 0$. The interior of region depicted represents all (b, M) such that signatories choose zero emissions.

48.2.2.2 Stage 1: The IEA Game

We now turn to the analysis of stable coalitions within the emissions game.

Proposition:

There exists $b_1(N)$, $b_2(N)$ such that:

- (a) If $b < b_1(N - 1)$, the unique stable IEA of the Stackelberg model with nonnegative emissions is the grand coalition (Proposition 3 in Rubio and Ulph (2006)).
- (b) If $b \in [b_2(N - 2), b_2(4) = N - 4]$, there exists an upper bound given by the smallest integer no less than n_3 that belongs to a self-enforcing IEA. This upper bound decreases with b (Proposition 4 in Rubio and Ulph (2006)).

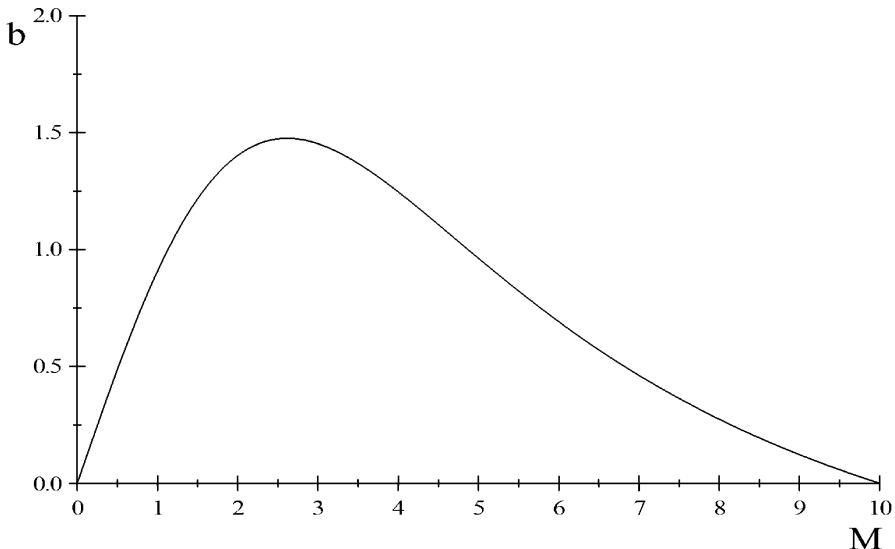


Fig. 48.2 Contour plot of $h(b, M)$

- (c) If $N > 5$ and $b > N - 4$, the maximum level of cooperation that can be achieved by a self-enforcing IEA is three (Proposition 5 in Rubio and Ulph (2006)).
- (d) If b is large enough, the equilibrium is interior and the largest size of a stable coalition is 2.

From the result above, one can conjecture that the size of the largest stable IEA is a decreasing function of b .

Thus, it is possible to get the grand coalition as a stable IEA, and this occurs when b is small enough. The possibility of a stable grand coalition is due to the leadership advantage of the IEA. Emissions are strategic substitutes (i.e., best responses are downward sloping). This in itself gives an incentive for a leader to increase its quantities relative to the case where it moves simultaneously with the non-signatories. The instability in the scenario where countries move simultaneously is due to the reaction of the outsider who increases its emissions following the creation of the IEA of $(N - 1)$ members. When the IEA is a leader, it decreases its overall level of emissions by a smaller amount than if it were not a leader (in which case, it possibly increases its emissions), and therefore, the outsider's increase in emissions is smaller than under the simultaneous-move game (where an outsider possibly decreases its emissions). This moderate reaction of the outsider is the reason why a grand coalition can be stable under a leadership model.

The externality of pollution induces the IEA (the leader) to reduce its emissions relative to the Nash equilibrium. In fact it can be shown that under the Nash equilibrium, an IEA may well end up with higher its emissions, resulting in a decrease of non-signatory emissions (possibly to a zero level). This is more likely to happen when b is small which explains the sustainability of the grand coalitions for small values of b . It is important to note that in a Nash equilibrium, the payoff of

non-signatories is always larger than that of signatories and in a Stackelberg equilibrium, this is no longer necessarily true. Interestingly, the emissions game, the range of parameters (b, M) under which an IEA is sustainable, corresponds to the range where the gains from cooperation are the largest.

48.3 Dynamic Games: Some Concepts

Natural resource and environmental problems usually involve interactions in a changing physical environment. Therefore, dynamic games are well suited for the analysis of many resources and environmental problems. (For a recent comprehensive survey of dynamic games in the economics of pollution, see Jorgensen, Martin-Herran, and Zaccour (2010). Long (2011) surveys dynamic games in natural resources.) Dynamic games are also called state-space games. In a state-space game, the environment is represented by a vector of state variables, which directly or indirectly affect the payoffs of agents. Agents influence the evolution of the state variables by using their control variables.

A dynamic game can be formulated in discrete time or in continuous time (see Long (2010, 2011) for surveys of models of both types). A dynamic game normally displays the following properties. Players receive a flow of benefits every period (or at every point of time). The overall payoff for a player is the sum (or integral) of his discounted flow of benefits over the time horizon. The benefit flow that a player receives in a period may depend on the current actions taken and on the “state of the system” in that period, as represented by the state variables. The state of the system changes over time, depending on the actions of the players. A difference equation or a differential equation describes the rate of change of each state variable.

The term “differential games” is broadly interpreted to include both dynamic games in continuous time and those in discrete time, where the evolution of each state variable is described by a difference equation.

Below is a description of a differential game in continuous time (see Dockner et al. (2000) for a more precise formulation). Time is represented by t . The game starts at time zero and ends at time T . There are n state variables, denoted by x_i where $i = 1, 2, \dots, n$. The vector of state variables is $\mathbf{x} = (x_1, x_2, \dots, x_n) \in X \subseteq \mathbb{R}^n$. The set $S \equiv X \times [0, T]$ is called the state-date space. An element (\mathbf{x}, t) is called a (state, date) pair. The number of player is an integer N . Player j has a vector of m control variables, denoted by u_j . Assume that $u_j(t) \in U_j \subseteq \mathbb{R}^m$. We call U_j player j ’s control space. Define $U \equiv \prod_j U_j$.

The evolution of the system is described by a system of s differential equations,

$$\dot{x}_i(t) = F_i(\mathbf{x}(t), \mathbf{u}_1(t), \mathbf{u}_2(t), \dots, \mathbf{u}_N(t), t), \quad i = 1, 2, \dots, n$$

where $x_i(0) = x_{i0}$ is given. In vector notation,

$$\dot{\mathbf{x}}(t) = \mathbf{F}(\mathbf{x}(t), \mathbf{u}_1(t), \mathbf{u}_2(t), \dots, \mathbf{u}_N(t), t)$$

Player j 's instantaneous flow of benefits at time t is

$$b_j(t) = B_j(\mathbf{x}(t), \mathbf{u}_1(t), \mathbf{u}_2(t), \dots, \mathbf{u}_N(t), t)$$

The time argument t will be suppressed when there is no risk of confusion. The overall payoff of player j is

$$\int_0^T e^{-r_j t} B_j(\mathbf{x}, \mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_N, t) dt + e^{-r_j T} S_j(\mathbf{x}_T, T)$$

where $S_j(x_T, T)$ is called the “salvage function” and $r_j \geq 0$ is the discount rate of player j .

A player can be a firm, or a government, or an individual, etc. Each player j maximizes its overall payoff. In order to do this, it must have some ideas about what other players are doing. A Nash equilibrium is a strategy profile such that each player's strategy maximizes its own overall payoff given what is predicted for the other players. (We focus on the case of simultaneous-move games because of space limitations. Games where agents play sequentially are called Stackelberg games (see, e.g., Benchekroun and Long (2011) Sect. 48.4.2 for more details on Stackelberg dynamic games in natural resource and environmental economics).) Such prediction depends on what strategy space each player is restricted to.

Consider two types of strategies: path strategies and Markovian decision-rule strategies (or feedback strategies). A path strategy (or open-loop strategy) p_j is a function that determines player j 's actions at each time t as a function of t and of the parameters of the model, including the initial stocks, but this function does not include the current value of the state variables. It is as if each player makes a commitment right at beginning of the game never to deviate from its planned time path of actions. Let P_j be the set of open-loop strategies that are available to player j . Let $P \equiv \prod_j P_j$. Once all players have chosen their open-loop strategies, the evolution of the state variables is described by

$$\dot{x}_i(t) = F_i(\mathbf{x}(t), \mathbf{p}_1(t), \mathbf{p}_2(t), \dots, \mathbf{p}_N(t), t), \quad i = 1, 2, \dots, n, \quad x_i(0) = x_{i0}$$

or, in vector notation,

$$\dot{\mathbf{x}}(t) = \mathbf{F}(\mathbf{x}(t), \mathbf{p}(t), t), \quad \mathbf{x}(0) = \mathbf{x}_0$$

where

$$\mathbf{p}(t) \equiv (\mathbf{p}_1(t), \mathbf{p}_2(t), \dots, \mathbf{p}_N(t))$$

Assume this equation has a unique solution $x^*(t)$. The overall payoff for player j is then

$$W_j(\mathbf{x}_0, \mathbf{p}) = \int_0^T e^{-r_j t} B_j(\mathbf{x}^*(t), \mathbf{p}(t), t) dt + e^{-r_j T} S_j(\mathbf{x}_T^*, T)$$

Define an open-loop Nash equilibrium (OLNE) as a strategy profile $\hat{\mathbf{p}} = (\hat{\mathbf{p}}_1, \hat{\mathbf{p}}_2, \dots, \hat{\mathbf{p}}_N) \in P$ such that no player can make itself better off by choosing a different open-loop strategy, i.e.,

$$W_j(\mathbf{x}_0, \hat{\mathbf{p}}) \geq W_j(\mathbf{x}_0, \mathbf{p}_j, \hat{\mathbf{p}}_{-j}) \text{ for all } j$$

To find an open-loop Nash equilibrium, one uses the maximum principle to derive the necessary conditions of each player's optimal control problem, taking as given the time path of the vector of control variables of other players. Then one finds a fixed point $\hat{\mathbf{p}}$ such that all the necessary conditions for all players are satisfied. Next one verifies that the sufficient conditions are satisfied at that fixed point.

One of the main advantages of the concept of open-loop Nash equilibrium is that such an equilibrium is relatively easy to find. Open-loop Nash equilibria are also attractive because they are time consistent. To see this, suppose the game is played and everyone has followed its Nash equilibrium strategy. Suppose at some time $t_1 > 0$, when the state vector takes on the value $x(t_1)$ as anticipated, player j asks itself whether it can make itself better off by switching to a different strategy. Clearly, the answer is no, because its original choice of strategy obeys Bellman's principle of optimality. (See, e.g., Leonard and Long 1992, Chap. 5 for a brief introduction to the principle of optimality.)

On the other hand, if by mistake some players have deviated from its planned course of action, so that the stock size $x(t_1)$ is different from what was anticipated at time zero, then at t_1 players will in general find that they would be better off by switching to another strategy. Therefore, open-loop Nash equilibria are not robust to "trembling hand" deviations (Selten 1975). One may say that open-loop Nash equilibria are not "subgame perfect" (even though the concept of a subgame is problematic in continuous time). For this reason, let us turn to the concept of Markov-perfect Nash equilibrium which overcomes this problem.

Define a Markovian decision-rule strategy (or simply Markovian strategy for short) as a function that determines at each (state, date) pair what action to take. Let ϕ_j be player j 's Markovian strategy, then

$$\mathbf{u}_j(t) = \phi_j(\mathbf{x}(t), t)$$

Let Q_j be the set of Markovian strategies that are available to player j . Let $Q \equiv \Pi_j Q_j$. Once all players have chosen their Markovian strategies, the evolution of the state variables is described by

$$\dot{x}_i(t) = F_i(\mathbf{x}(t), \phi_1(\mathbf{x}(t), t), \dots, \phi_N(\mathbf{x}(t), t), t), \quad i = 1, 2, \dots, n, \quad x_i(0) = x_{i0}$$

or, in vector notation,

$$\dot{\mathbf{x}}(t) = \mathbf{F}(\mathbf{x}(t), \phi(t), t), \quad \mathbf{x}(0) = \mathbf{x}_0$$

where

$$\phi(t) \equiv (\phi_1(t), \phi_2(t), \dots, \phi_N(t))$$

Assume this differential equation has a unique solution for any initial condition (x_{t_1}, t_1) . Define the performance index for player i at the (state, date) pair (x, t) by

$$J_j(\mathbf{x}, t, \phi) = \int_t^T e^{-r_j(\tau-t)} B_j(\mathbf{x}(\tau), \phi(\mathbf{x}(\tau), \tau), \tau) d\tau + e^{-r_j(T-t)} S_j(\mathbf{x}_T, T)$$

We define a Markov-perfect Nash equilibrium (also called a feedback Nash equilibrium) as a strategy profile $\hat{\phi} = (\hat{\phi}_1, \hat{\phi}_2, \dots, \hat{\phi}_N) \in Q$ such that, at any (state, date) pair $(x, t) \in X \times [0, T]$, no player can make itself better off by choosing a different strategy, i.e.,

$$J_j(\mathbf{x}, t, \hat{\phi}) \geq J_j(\mathbf{x}, t, \phi_j, \hat{\phi}_{-j}) \text{ for all } j$$

It is important to stress the requirement that this inequality be satisfied for all possible (state, date) pair $(x, t) \in X \cup [0, T]$, not just for the initial pair at time zero $(x_0, 0)$. As Reinganum and Stokey (1985) point out, a decision-rule Nash equilibrium for a given $(x_0, 0)$ is not necessarily Markov perfect. To be Markov perfect, a Nash equilibrium in decision rules must satisfy the additional property that the continuation of the given decision rules constitutes a Nash equilibrium when viewed from any future (date, state) pair. Dockner et al. (2000, example 4.2) give an example of a Nash equilibrium in decision rules that fails to be Markov perfect.

To find a Markov-perfect Nash equilibrium (MPNE), the usual method is to make use of the Hamilton-Jacobi-Bellman (HJB) equations that the value function of each player must satisfy. The HJB equation for player j is

$$rV_j(\mathbf{x}, t) - \frac{\partial V_j(\mathbf{x}, t)}{\partial t} = \max_{\mathbf{u}_j} \left\{ B_j(\mathbf{x}, \mathbf{u}_j, \hat{\phi}_{-j}(\mathbf{x}), t) + \frac{\partial V_j(\mathbf{x}, t)}{\partial \mathbf{x}} \mathbf{F}(\mathbf{x}, \mathbf{u}_j, \hat{\phi}_{-j}(\mathbf{x}), t) \right\}$$

with the terminal condition

$$V_j(\mathbf{x}, T) = S_j(\mathbf{x}, T)$$

If T is infinite, the above terminal condition is replaced by

$$\lim_{t \rightarrow \infty} e^{-rt} V_j(\mathbf{x}(t), t) = 0$$

It is worth noting that OLNE and MPNE can be thought of as based on two alternative assumptions about the ability of players to precommit. In an OLNE, players commit to a whole time path of actions. In an MPNE, players cannot precommit at all. Reinganum and Stokey (1985) argue that in some cases, players may be able to commit to actions in the near future (e.g., by forward contracts), but not to actions in the distant future. They develop a simple model where a game begins at time 0 and ends at a fixed time T , and there are k periods of equal lengths δ , where $k\delta = T$. At the beginning of each period, agents can commit to a path of action during that period. The special case where $k = 1$ corresponds to the open-loop formulation, and OLNE is then the appropriate equilibrium concept. At the other extreme, where $\delta \rightarrow 0$, the appropriate equilibrium concept is MPNE.

The choice of equilibrium concepts is to some extent dependent on tractability. The relative ease of finding an OLNE is one of its attractive features. For some examples of OLNE in the economics of natural resources, see Gaudet and Long (1994) and Benchekroun et al. (2009).

48.4 Transboundary Stock Pollutants

48.4.1 A Benchmark Model

Following Long (1992) and Ploeg and Zeeuw (1992), let us consider a world consisting of two countries. Let $Q_i(t)$ be country i 's output at date t . Assume that emissions are proportional to output, $E_i(t) = Q_i(t)$. Let $P(t)$ denote the stock of pollution. Assume

$$\dot{P}(t) = E_1(t) + E_2(t) - \delta P(t) \quad (48.2)$$

where $\delta > 0$ is the decay rate. The pollution damage suffered by country i at time t is $\frac{cP^2}{2}$. The net utility of country i is

$$U_i(t) = AE_i(t) - \frac{1}{2} (Q_i(t))^2 - \frac{c}{2} (P(t))^2$$

and its social welfare is

$$W_i = \int_0^\infty e^{-\rho t} U_i(t) dt$$

where $\rho > 0$ is the rate of discount.

Let us find the open-loop Nash equilibrium of this model. Since countries use path strategies in the open-loop formulation, let us suppose that country i believes that country j 's emission strategy is $E_j(t) = g_j^{OL}(t)$. Then it seeks to solve the following optimal control problem:

$$\max_{E_i(\cdot)} \int_0^\infty e^{-\rho t} \left[AE_i(t) - \frac{1}{2} (E_i(t))^2 - \frac{c}{2} (P(t))^2 \right] dt \quad (48.3)$$

subject to

$$\dot{P}(t) = E_i(t) + g_j^{OL}(t) - \delta P(t), \quad P(0) = P_0 \quad (48.4)$$

Applying the maximum principle, we obtain the necessary conditions

$$\begin{aligned} -(\dot{E}_i - \rho(E_i - A_i)) &= -c_i P - \delta(E_i - A_i) \\ \dot{P} &= E_i + g_j^{OL} - \delta P, \quad P(0) = P_0 \end{aligned}$$

and the transversality condition is

$$\lim_{t \rightarrow \infty} e^{-\rho t} (E_i(t) - A_i) P(t) = 0$$

Since the two countries are identical, we obtain the following system of two differential equations

$$\dot{E} = cP + (\rho + \delta)(E - A) \quad (48.5)$$

$$\dot{P} = 2E - \delta P, \quad P(0) = P_0 \quad (48.6)$$

with the transversality condition

$$\lim_{t \rightarrow \infty} e^{-\rho t} (E(t) - A) P(t) = 0 \quad (48.7)$$

There is a unique steady-state pair $(P_\infty^{OL}, E_\infty^{OL})$ where

$$P_\infty^{OL} = \frac{2A(\delta + \rho)}{2c + \delta(\delta + \rho)} \quad (48.8)$$

$$E_\infty^{OL} = \frac{A\delta(\delta + \rho)}{2c + \delta(\delta + \rho)} = \frac{\delta P_\infty^{OL}}{2} \quad (48.9)$$

Comparing with the case where the two countries cooperate and maximize the sum of their welfare, we see that the steady-state stock of pollution P_∞^{OL} is too high.

What happens if countries use feedback strategies? Suppose country i believes that country j employs a feedback emission strategy, $E_j(t) = g_j^{FB}(P(t))$, so that its rate of emissions at t is conditioned on the currently observed level $P(t)$. Then country i maximizes

$$\max_{E_i(\cdot)} \int_0^\infty e^{-\rho t} \left[AE_i(t) - \frac{1}{2} (E_i(t))^2 - \frac{c}{2} (P(t))^2 \right] dt \quad (48.10)$$

subject to

$$\dot{P}(t) = E_i(t) + g_j^{FB}(S(t)) - \delta P(t), \quad P(0) = P_0 \quad (48.11)$$

Realizing that $g_j^{FB}(P)$ is a function of the pollution stock, country i knows that it can indirectly manipulate country j 's emissions at t by influencing the evolution of P . This strategic consideration was absent in the open-loop case.

To find the feedback Nash equilibria of this game, we make use of the Hamilton-Jacobi-Bellman (HJB) equations. The HJB equation for country i is

$$\rho V_i(P) = \max_{E_i} \left[AE_i - \frac{1}{2} E_i^2 - \frac{c}{2} P^2 + V'_i(P)(E_i + E_j(P) - \delta P) \right]$$

where $E_j(P)$ is country j 's feedback strategy and $V_i(P)$ is country i 's value function. The transversality condition is

$$\lim_{t \rightarrow \infty} e^{-\rho t} V_i(P(t)) = 0 \quad (48.12)$$

The first-order condition with respect to E_i is $E_i = A + V'_i(P)$. This equation gives $E_i = E_i(P)$, i.e., country i 's emissions depend only on P . Appealing to symmetry, we get the HJB equation

$$\rho V(P) = \frac{1}{2} \left[A^2 + 4AV' + 3(V')^2 \right] - \delta PV' - \frac{c}{2} P^2 \quad (48.13)$$

This equation and the transversality condition Eq. (48.12) identify the set of possible Markov-perfect Nash equilibria.

Let us conjecture that the value function is quadratic

$$V(P) = -\frac{\omega P^2}{2} - \pi P - \mu \quad (48.14)$$

Then $V'(P) = -\omega P - \pi$ and hence the feedback strategy is linear

$$E(P) = A - \pi - \omega P \quad (48.15)$$

It is plausible to expect that $\omega > 0$, i.e., a higher stock will make countries choose lower emissions, and $\pi > 0$, i.e., if $P = 0$, the marginal effect on welfare of an exogenous increase in P is negative.

Making use of Eq. (48.14) and Eq. (48.15), the HJB equation gives a quadratic equation of the form

$$\lambda_0 + \lambda_1 P + \lambda_2 P^2 = 0$$

where λ_0, λ_1 , and λ_2 are expressions involving the parameters δ, ρ, c and the coefficients ω, π, μ . Since this equation must hold for all P , it follows that $\lambda_i = 0$ for $i = 0, 1, 2$. Using these three conditions, we can solve for ω, π , and μ . We obtain

$$\omega = \frac{1}{3} \left[-\left(\delta + \frac{\rho}{2} \right) + \sqrt{\left(\delta + \frac{\rho}{2} \right)^2 + 3c} \right] \quad (48.16)$$

(To ensure convergence to a steady state, the positive root $\omega > 0$ is selected.) Next, compute π and μ as follows:

$$\begin{aligned} \pi &= \frac{2A\omega}{\delta + \rho + 3\omega} \\ \mu &= \frac{(A - \pi)}{2\rho} (3\omega - \delta - \rho) \equiv \mu_m \end{aligned}$$

The linear feedback strategy is

$$E = \frac{A(\delta + \rho + \omega)}{\delta + \rho + 2\omega} - \omega P$$

It follows that

$$\dot{P} = \frac{2A(\delta + \rho + \omega)}{\delta + \rho + 2\omega} - (2\omega + \delta)P \quad (48.17)$$

For P to converge to a steady state, it is necessary that $2\omega + \delta > 0$. This inequality is satisfied if and only if the positive root for ω is selected. The steady-state pollution stock under the MPNE with linear feedback strategies is

$$P_\infty^{FB} = \frac{2A(\delta + \rho + \omega)}{(\delta + \rho + 2\omega)(2\omega + \delta)} \quad (48.18)$$

Clearly, the OLNE steady-state pollution stock P_∞^{OL} is lower than the MPNE steady state P_∞^{FB} . This result is dependent on the fact that we have focused on a quadratic functional form for the value function $V_i(P)$. Dockner and Long (1993) show that there are other value functions that satisfy the HJB equation. These value functions result in nonlinear emission strategies. In fact there is a continuum of nonlinear strategies, and some of them outperform the OLNE in the sense that both countries would be better off under such strategies. When there are multiple equilibria, it is not clear which one is likely to prevail.

48.4.2 Centralized Versus Regional Control of Pollution

List and Mason (2001) consider an asymmetric version of the transboundary pollution model of Dockner and Long (1993) in the case of two regions and where pollution management can be centralized, i.e., dictated by a federal authority such as the EPA for the case of the USA, or CEA in Canada, or decentralized, i.e., regulation is chosen by local states or provinces. In their model, there are two regions that have different parameter values for the regional damage function and production function:

$$U_1(t) = AE_1(t) - \frac{1}{2} (Q_1(t))^2 - \frac{c}{2} (P(t))^2$$

and

$$U_2(t) = \alpha AE_2(t) - \frac{1}{2} (Q_2(t))^2 - \beta \frac{c}{2} (P(t))^2$$

where α, β characterize differences between the two regions in vulnerability to flow and stock pollution as well as differences in abatement costs. An alternative interpretation is that differences in instantaneous utilities are the result of population differences.

They characterize the equilibrium obtained when a central authority, whose objective is to maximize the sum of the two regions discounted sum of welfare, sets the environmental policy, assuming it is constrained by the constitution to set uniform environmental policies in both regions. Given the asymmetry of the two regions, the central authority cannot achieve a first best. They show that if $\beta = 0$, i.e., one region is not affected by the stock of pollution and $\alpha = 1$, then welfare values under central control exceed those under decentralized control. However when $\beta = 0$ and α is large enough then, for small values of the stock of pollution, the present value of combined payoffs for the two regions is larger under decentralized than centralized control. The larger the asymmetry between the two regions, the larger the cost of implementing a central authority's plan under the constraint of uniform regulation. When the asymmetry is large enough, the distortion introduced by the constraint outweighs the gains from the elimination of free riding (under a central authority). This result can be extended to the case where $\beta > 0$ using a continuity argument. When $\beta = 0$, it is shown that decentralization always results in higher rates of emissions and a larger steady-state stock of pollution than under central management.

List and Mason (1999) examine whether environmental regulations should be carried out locally or centrally; i.e., by a central authority or by local regulators. Localities are assumed to have superior information or more leniency to adopt new environmental regulations. They consider the case of several pollutants. Consider two regions (states or provinces) indexed by $i = 1, 2$. In each region i production generates two flow of emissions, one local and one transboundary, denoted by F_i

and E_i . These flows of emissions accumulate and form stocks of pollutants: a local stock pollutant denoted by Z_i and a transboundary pollution stock by P . The evolution of the stocks is given by

$$\dot{Z}_i = F_i - lZ_i$$

and

$$\dot{P} = E_1 + E_2 - kP$$

where $k, l > 0$ are nature's purification rates for the transboundary and the local pollutant, respectively.

The instantaneous utility U_i of country i is given by

$$U_i = AE_i + BF_i - \frac{1}{2}E_i^2 - \frac{1}{2}E_i^2 - \frac{X}{2}Z_i^2 - \frac{S}{2}P_i^2 - \rho Z_i P$$

where X and S are positive damage parameters. The parameter ρ captures the interaction between the local and the transboundary pollutants. When the interaction reduces damages, we have $\rho < 0$, and when $\rho > 0$, the two pollutants have synergistic negative effects. It is assumed that local authorities know the value of ρ , whereas the central authority ignores the true value of ρ and uses a value of $\rho = 0$ when choosing the optimal emission policy under a centralized system.

They examine when local regulation dominates a central system in the case of carbon dioxide (the transboundary pollutant) and sulfur (local pollutants) and use parameter values based on empirical evidence. They show that there exist $\rho^+ > 0$ and $\rho^- < 0$ such that the benefits from local control can more than offset the benefits from central control if synergistic effects are such that $\rho > \rho^+$ or $\rho < \rho^-$.

48.5 Provision of Clean Air and Interregional Mobility of Capital

Regional governments may impose capital income tax to finance the provision of public goods such as clean air. However, a tax imposed on the earning of a factor of production will encourage that factor to move to another jurisdiction where the tax rate is lower. (With the exception of completely immobile factors, such as land and mineral resources.) Stigler (1965) points out that if all factors are mobile, redistributive taxation in a multi-jurisdiction world is practically infeasible. Zodrow and Mieszkowski (1986) show that if regional governments compete in source-based capital income tax rates, there will be a race to the bottom, leading to the underprovision of local public goods. The theoretical literature has identified size differences as a factor for explaining why different jurisdictions are affected asymmetrically by tax competition (Bucovetsky 1991; Wilson 1991). For generalization to two tax instruments, see Bucovetsky and Wilson (1991). Wang (1999) assumes sequential moves: the bigger region is the Stackelberg leader.

A number of two-period models have been developed to investigate the implications of simple dynamic games between the owners of partially mobile factors of production on the one hand and a local government that tries to redistribute income in favor of some group, on the other hand. Lee (1997) shows that if capital movements involve adjustment costs, there will be a wedge between the internal rate of return and the external one. Jenson and Thomas (1991) model a game between two governments that use debt policies to influence the intertemporal structure of taxation. Huizinga and Nielsen (1997) formulate a two-period model in which even though capital is perfectly mobile, foreign capitalists in effect earn rents from local immobile resources.

Wildasin (2002, 2008) presents a continuous-time, infinite-horizon model in which infinitely lived agents react to changes in taxation by moving resources across jurisdictions. Adjustment costs are explicitly taken into account. The author focuses on the case of a once-over tax change and does not deal with optimal time-varying tax rates. Instead, the analysis emphasizes the costly adjustment process and draws on the adjustment cost literature in macroeconomics (Turnovsky 2000). The main point is that capital earns quasi rent which can be taxed away, but such quasi rents erode with time. The optimal capital income tax rate depends crucially on the degree of capital mobility. Wildasin does not model a dynamic game involving the competition among jurisdictions to attract mobile resources.

A truly dynamic game model is that of Koethenbuerger and Lockwood (2010). The authors consider an infinite-horizon dynamic version of the model of Zodrow and Mieszkowski (1986). There are n regions, with one firm in each region. Each region is subject to a stochastic output shock. These shocks imply that households would like to diversify their portfolios, and this dampens the tax competition among regional governments. Under logarithmic utility, they show that the Nash equilibrium path of capital income tax rate is time invariant. This constant tax rate is increasing in the preference parameter for the public good, the rate of discount, and the volatility of the output shock. There exists a critical threshold \hat{n} such that the equilibrium tax rate is increasing in n if $n < \hat{n}$ and decreasing in n if $n > \hat{n}$. As n tends to infinity, the equilibrium tax rate tends to zero, which is an inefficient outcome.

48.6 Conclusions

Both static and dynamic games have been successfully employed to shed light on many resource and environmental issues involving strategic interactions among a number of players. The insights generated by game theoretic models can potentially be used to help design mechanisms for improving economic efficiency. In particular, empirical models are useful tools for policy making. Conversely, issues in resource and environmental economics have provided opportunities for researchers to sharpen their tools and to develop new concepts and techniques for dealing with emerging issues. Because of space limitations, we have covered noncooperative games only. We have omitted empirical models of dynamic

games in resource and environmental economics that use real world data to calibrate parameters of demand and cost functions. We have also omitted games with asymmetric information. We refer the reader to the recent surveys in Long (2010, 2011) and Jorgensen et al. (2010).

References

- Barrett S (1994) Self-enforcing international environmental agreements. *Oxf Econ Pap* 46:878–894
- Benchekroun H, Long NV (2011) Static and dynamic games in environmental and resource economics. In: Batabayal A, PeterNijkamp P (eds) *Research tools in natural resource and environmental economics*. World Scientific, Hackensack
- Benchekroun H, Halsena A, Withagen C (2009) On nonrenewable resource oligopolies: the asymmetric case. *J Econ Dyn Control* 33:1867–1879
- Bucovetsky S (1991) Asymmetric tax competition. *J Urban Econ* 30(2):167–181
- Bucovetsky S, Wilson J (1991) Tax competition with two tax instruments. *Reg Sci Urban Econ* 21(3):333–350
- Dockner EJ, Jorgensen S, Long NV, Sorger G (2000) Differential games in economics and management science. Cambridge University Press, UK
- Dockner E, Long NV (1993) International pollution control: cooperative versus non-cooperative strategies. *J Environ Econ Manag* 25:13–29
- Gaudet G, Long NV (1994) On the effects of the distribution of initial endowments in a non-renewable resource duopoly. *J Econ Dyn Control* 18:1189–1198
- Huizinga H, Nielsen SB (1997) Capital income and profit taxation with foreign ownership of firms. *J Int Econ* 42:149–165
- Jenson R, Thomas EF (1991) Debt in a model of tax competition. *Reg Sci Urban Econ* 21:371–392
- Jorgensen S, Martin-Herran G, Zaccour G (2010) Dynamic games in the economics and management of pollution. *Environ Model Assess.* doi:10.1007/s10666-010-9221-7
- Koethenbuerger M, Lockwood B (2010) Does tax competition promote growth? *J Econ Dyn Control* 34(2):191–206
- Lee K (1997) Tax competition with imperfectly mobile capital. *J Urban Econ* 42:222–242
- Leonard D, Long NV (1992) Optimal control theory and static optimization in economics. Cambridge University Press, New York/Cambridge, UK
- List JA, Mason CF (1999) Spatial aspects of pollution control when pollutants have synergistic effects: evidence from a differential game with asymmetric information. *Ann Reg Sci* 33(4):439–452
- List JA, Mason CF (2001) Optimal Institutional arrangements for transboundary pollutants in a second-best world: evidence from a differential game with asymmetric players. *J Environ Econ Manag* 42(3):277–296
- Long NV (1992) Pollution control: a differential game approach. *Ann Oper Res* 37:283–296
- Long NV (2010) A survey of dynamic games in economics. World Scientific, Singapore
- Long NV (2011) Dynamic games in the economics of natural resources: a survey. *Dyn Games Appl* 1(1):115–148
- Reinganum JF, Stokey NL (1985) Oligopoly extraction of a common property natural resource: the importance of period of commitment in dynamic games. *Int Econ Rev* 26:161–173
- Rubio S, Ulph A (2006) Self-enforcing international environmental agreements revisited. *Oxf Econ Pap* 58(2):233–263
- Selten R (1975) Reexamination of perfectness concepts for equilibrium points in extensive form games. *Int J Game Theory* 4(1):25–55
- Stigler G (1965) The tenable range of functions of local government. In: Phelps ES (ed) *Private wants and public needs*. Norton, New York, pp 167–176
- Turnovsky SJ (2000) Methods of macroeconomic dynamics. MIT Press, Cambridge, MA

- van der Ploeg F, de Zeeuw AJ (1992) International aspects of pollution control. *Environ Res Econ* 2:117–139
- Wang Y-Q (1999) Taxes under fiscal competition: stackelberg equilibrium and optimality. *Am Econ Rev* 89(4):947–981
- Wildasin DE (2002) Fiscal competition in space and time? *J Public Econ* 87:2571–2588
- Wildasin DE (2008) Fiscal competition for imperfectly mobile labor and capital: a comparative dynamic analysis. CESifo working paper 2808, University of Munich
- Wilson J (1991) Tax competition with interregional difference in factor endowments. *Reg Sci Urban Econ* 21(3):423–451
- Zodrow GR, Mieszkowski P (1986) Pigou, tiebout, property taxation, and the under-provision of local public goods. *J Urban Econ* 19:356–370

Economic Valuation: Concepts and Empirical Methods

49

John B. Loomis

Contents

49.1	Introduction	974
49.2	Benefit Measures	975
49.2.1	Use Values	975
49.2.2	Nonuse or <i>Passive Use Values</i>	975
49.3	Overview of Methods and How They Relate to Values	976
49.4	Hedonic Property Method	977
49.4.1	Economic Theory Underlying the Hedonic Property Method	977
49.4.2	Data Requirements	978
49.4.3	Econometric Modeling Including Spatial Dimensions	979
49.5	Travel Cost Models	979
49.5.1	<i>Trip Frequency Models</i> of Recreation Demand	980
49.5.2	Multisite Selection Models	982
49.5.3	Data Requirements for Travel Cost Models	982
49.6	Stated Preference Models	983
49.6.1	Contingent Valuation Method	983
49.6.2	Choice Experiments	985
49.6.3	The Issue of Bias in Stated Preference Surveys	987
49.7	Combining Stated and Revealed Preference Methods and Data	988
49.8	Benefit Transfer	988
49.9	Conclusions	990
	References	991

J.B. Loomis

Department of Agricultural and Resource Economics, Colorado State University, Fort Collins,
CO, USA
e-mail: jloomis@lamar.colostate.edu

Abstract

Commensurate valuation of market and nonmarket public goods allows for a more valid benefit-cost analysis. Economic methods for valuing nonmarket public goods include actual behavior-based revealed preference methods such as the hedonic property method for urban-suburban public goods and travel cost-based models for outdoor recreation. For valuing proposed public goods for which there is no current behavior or valuing the existence or passive use values of public goods, economists can rely upon stated preference methods. While there is skepticism among some economists for relying upon what people say they will pay rather than what their actual behavior suggests they will pay, there is general acceptance of stated preference methods. These stated preference methods include the well-known contingent valuation method and choice experiments (sometimes called conjoint analysis). Lastly, in situations where there is neither time nor money to conduct an original revealed or stated preference study, economists typically rely upon benefit transfers from existing revealed preference and stated preference studies to provide rough estimates of the values of public goods such as water quality, air quality, wetlands, recreation, and endangered species.

49.1 Introduction

One of the long-standing deviations from economic efficiency of even a perfectly competitive market with no subsidies to producers or consumers is that of negative externalities and provision of *public goods*. In the face of these market failures, government intervention has the potential to improve economic efficiency by imposing pollution taxes or tradeable permits to internalize the negative externalities into prices of the goods associated with pollution. Further, government has the *potential* to improve economic efficiency by supplying or financing the supply of optimal amounts of the public good.

However, the emphasis here is on the potential to improve economic efficiency through government action. For this potential to be realized, the level of the pollution taxes must be set equal to the marginal environmental cost at the socially optimum level of output. Thus, to achieve this optimum requires having an estimate of the marginal environmental cost of pollution or, alternatively, the marginal benefits of improving environmental quality (e.g., air quality, water quality). The same is true of public goods: the government has to determine the marginal benefits of these public goods to society so as to compare to the cost of producing alternative levels of the public goods to determine an optimum.

Benefit-cost analysis is a technique used by government to determine if the benefits of increased environmental quality or public goods are worth the cost. One of the greatest challenges of benefit-cost analysis is estimating nonmarket benefits of regulations imposed on industry to internalize negative externalities (e.g., installation of pollution control devices) or government supply of public goods (e.g., preservation of remote wilderness areas).

This chapter is devoted to a review of environmental valuation methods frequently used by a wide variety of economists (i.e., academic, government, consultants) to estimate the economic benefits of improving environmental quality and public goods. The conceptual foundation of all environmental valuation methods is reviewed first. This is followed by a discussion of actual behavior-based environmental valuation methods. These methods are usually referred to as revealed preference methods and include the hedonic property method and the travel cost method. This section is followed by a review of stated preferences methods including the contingent valuation method and choice experiments. The next to the last section discusses how revealed preference and stated methods can be combined to provide more robust environmental valuations. Finally, “shortcut” methods called benefit transfer are reviewed.

49.2 Benefit Measures

Value has many different meanings, and it is important for economists to be precise as to what they mean by economic value or benefits of environmental quality or public goods. The economic value or benefit received by a person for any good whether marketed or nonmarketed is the maximum amount they would pay for it. The term economists used for this is maximum *willingness to pay* (WTP). WTP is short hand for willingness and ability to pay. When estimated as the area under a consumer’s demand curve, it is usually referred to as consumer surplus. While there are many theoretical refinements to this measure, for an applied economist, consumer surplus is generally considered a reasonable approximation to these more theoretically correct concepts of consumer well-being.

It is worth noting that nothing has been said about jobs created by production of a public good as an economic efficiency benefit or jobs lost with environmental regulation as a cost. Except in times of unusually high and persistent unemployment, gains in jobs in one industrial sector are usually made up in another. Likewise, jobs lost in one geographic area are usually made up in another. Hence, jobs are considered transfers of economic activity from one industrial sector or geographic area to another. In other words, changes in jobs are not net gains or net losses to the economy as a whole and are usually excluded from an economic efficiency analysis such as benefit-cost analysis.

49.2.1 Use Values

For most market and nonmarket goods and services, the benefits are largely received by individuals who actually consume or directly use the good. The benefits of another hamburger or a new reservoir are primarily to the consumers who use it. In the reservoir example, use values would accrue to those who receive drinking water from the reservoir, receive flood protection, or water ski at the reservoir. The vast majority of benefits from a project or policy typically fall into the use category

as this is a very broad category. Use values also include the value of publicly provided recreation, scenic visibility at national parks, and commonly seen wildlife such as deer. Use values also relate to reduction in health damages from cleaning up hazardous waste sites and improving air and water quality. These use values are also measured by the users' maximum willingness to pay, so that there is consistency between valuation of market goods and nonmarket goods, i.e., the dollars are commensurate.

49.2.2 Nonuse or Passive Use Values

There are, however, unique natural resources such as Yellowstone National Park, rare/endangered species such as condors or panda bears from which people often receive benefits from just knowing these exist in the wild. This type of value is known as *existence value* (Krutilla 1967; Freeman 2003). Receiving this benefit does not require an on-site visit. Rather, there is an enjoyment from reflecting on the existence of the Arctic National Wildlife Refuge in Alaska undisturbed by oil and gas drilling. Likewise some people receive enjoyment and satisfaction that protection of these unique natural environments or species today will provide to future generations. This "bequest value" also does not require the current lived person to set foot in the area or personally view it.

The existence and bequest values are sometimes called nonuse values (Freeman 2003) or passive use values (US District Court of Appeals 1989; Arrow et al. 1993). These values have been the focus of natural resource damage assessment (e.g., damages from oil spills in remote areas of Alaska from the Exxon Valdez oil tanker spill – see Carson et al. 2003) and biodiversity (see Abdullah et al. (2011) for a review of these valuation studies).

Given that everyone can simultaneously enjoy the knowledge that a given unique natural environment exists, existence values have the characteristics of public goods. If valuing public goods were not difficult enough, these nonuse public goods are particularly challenging since there is little tie to a consumer's behavior. However, as discussed later in this chapter, economists have developed and implemented stated preference valuation methods that can measure the benefits of these special types of public goods. These passive use values are also measured by the maximum amount that people who benefit from these public goods would pay for them. This insures consistency between passive use values and use values and market values.

49.3 Overview of Methods and How They Relate to Values

There are two broad classes of valuation methods for nonmarket resources. Revealed preference methods refer to methods that indirectly infer WTP based on market transactions for other related goods. For example, estimating a demand curve for recreation based on the variation in visitors' travel costs. From the demand curve, visitors' WTP or consumer surplus can be calculated. The generic

label for this type of revealed preference method is *Travel Cost Model* because it relies on travel behavior and travel costs. Another revealed preference method is the *Hedonic Property Method*. This method disaggregates the price of a house purchased into the attributes of the house itself (e.g., bedrooms, bathrooms), the neighborhood (e.g., school quality), and the surrounding environment (e.g., distance to work, distance to an amenity or disamenity to be valued). Since houses with proximity to desirable environmental attributes are demanded by more households, this pushes up their prices. The price premium for a location close to an amenity such as open space or a park or good air quality can then be inferred.

In contrast, stated preference methods such as the *Contingent Valuation Method* or *Choice Experiments* rely upon what people say they would intend to pay if a certain scenario occurs. For example, how much more I would pay in trip cost for access to a recreation site with better water quality or how much more I would pay in taxes to protect an endangered species in a remote area. As will be discussed in more detail below, stated preference methods have the advantage of being quite flexible so it can measure both use and passive use values. Stated preference methods can also value a wide range of public goods including health, air quality, water quality, recreation, and endangered species. This flexibility comes at a price of potential hypothetical bias where respondents to the survey may state they will pay more for the public good than they would actually pay when they must hand over their own hard-earned money. Below we talk about what the literature finds with regard to when hypothetical bias is more likely to occur and what can be done to reduce it.

It is important to emphasize that all these estimation methods are just alternative tools for measuring WTP. They do it differently, but the measure of value is still the same. At the end of this chapter, we will also talk about how revealed preference and stated preference data can be combined to utilize the strengths of each method. But for now, we will discuss each method separately.

49.4 Hedonic Property Method

This revealed that preference technique has been applied to estimating house price differentials with natural hazards (e.g., earthquakes, floods, fires), environmental quality (air pollution, water pollution), and recreation access (e.g., open space, beaches). To understand how this versatile technique works, we will first review the theory underlying it, the data requirements, then the econometric estimation, and finally how WTP is calculated from the regression results.

49.4.1 Economic Theory Underlying the Hedonic Property Method

Competition for houses with desirable amenities pushes the prices of these houses up. Likewise, to entice home buyers to purchase homes with less desirable locations or disamenities, sellers must lower their prices. These premiums and discounts are intuitive but in order to develop valid estimates of WTP, there must be a close link

between the theoretical foundation and the empirical estimation. Further, any empirical model is based on a set of assumptions, which are often embedded in the theory. Below, we summarize the theory (see Taylor 2003 for more comprehensive discussion of the theory and empirical methods discussed below).

In the hedonic property method, the standard assumptions that consumers maximize utility and sellers maximize profits are employed. The consumer's utility function is Lancasterian in nature being specified in terms of the attributes of the house structure itself and its location. A stylized representation of the utility function is

$$U_i(X, A_s, A_n, A_e) \quad (49.1)$$

where U_i is utility of person i and X represents all other nonhousing goods and is sometimes referred to as a composite commodity. The A 's represent attributes of the housing structure itself (A_s) (e.g., bedrooms, baths), the neighborhood (A_n) (e.g., education levels), and the environment (A_e) (e.g., air quality, water quality). This utility function is maximized subject to the consumer's budget constraint (where the price of the composite commodity is normalized to 1). The consumer optimum is where

$$\partial Ph/\partial A_i = (\partial U/\partial A_i)/(\partial U/\partial X) \quad (49.2)$$

where Ph is the price of the house.

The interaction of the producers' minimum willingness to accept to supply attributes and consumers' maximum WTP for attributes results in an equilibrium price schedule for attributes A_s , A_n , and A_e . In an equilibrium between the producer and consumers, $\partial Ph/\partial A_i$ is the marginal WTP for small changes in A_i .

From the theory comes an estimable hedonic price function. In Eq. (49.3), we present an illustrative form of it:

$$Ph = \beta_0 + \beta_1(HA) + \beta_2(SQ) + \beta_3(NInc) + \beta_4(DWork) + \beta_5(EQ) \quad (49.3)$$

where Ph is the price of the house; HA is housing structure size; SQ is school quality, e.g., graduation rates; $NInc$ is neighborhood income – often proxied by census tract or zip code; $DWork$ is distance to employment centers; and EQ , e.g., air quality (parts per million of key pollutants), is distance to open space or a disamenity like a landfill.

The implicit price or marginal WTP for a small change in any attribute of the house structure, neighborhood, or environmental quality is simply the regression slope coefficient if the hedonic price function is linear. If the house price function is nonlinear, as it typically is, then the contribution of each additional unit of attribute to the house price is also related to the absolute level of house price. In this case, the formula for marginal WTP is slightly more complicated. Taylor (2003) provides formulas for the implicit price function for a variety of nonlinear functional forms.

Since the implicit price function is for a marginal change in attribute levels, it will overstate the benefits of a large increase in attributes but understate the loss of large changes in attributes. In order to accurately estimate the benefits for large gains or losses in attributes, a second-stage hedonic demand for the specific attribute must be estimated. Discussion of this is beyond the scope and space available in this chapter so the interested reader should see Taylor (2003) for more details.

49.4.2 Data Requirements

The data required for this method is of course quite detailed. The analyst needs house sale prices, characteristics of the home, characteristics of the neighborhood, and characteristics of the environment. This requires obtaining at least three different data sources. House sale prices and house characteristics are often available from county tax assessors' offices or from third-party real estate services. Characteristics of the neighborhood such as income, ethnicity, and average age are often found in block level data available from a government's population census office or sold by third-party vendors. Data on environmental quality of the neighborhood is often obtained from some form of monitoring station or field data. Location of houses relative to the amenity or the disamenity must often be calculated using Geographic Information System software. This requires that housing data be "georeferenced" in some form whether street address or coordinates. Needless to say that assembling the data can be time consuming, but no more so than the other methods we will review.

49.4.3 Econometric Modeling Including Spatial Dimensions

Since the implicit prices are essentially the regression coefficients, an econometric model must be estimated using the data assembled above. Historically, nearly all hedonic price functions were estimated using ordinary least squares regression in one form or another. Recently, there have been concerns that there may be spatial dependence of prices between houses located in close proximity to one another (e.g., same neighborhoods). This dependence may be due to real estate agents and appraisers' use of "comparable houses" when determining fair market value or appraised value for houses. It may also be due to there being some unobservable (to the analyst) characteristic of a particular neighborhood shared by houses in that neighborhood. Since this characteristic is unobservable to the analyst, it is an omitted variable in the regression equation. In the last few years, spatial econometric methods have been developed to address these problems (Anselin 1988). At present, some studies show that using these more advanced methods may result in more accurate estimates of the implicit prices, but in other cases, there is little difference (Mueller and Loomis 2008). The interested reader should see Anselin (1988) for more details.

49.5 Travel Cost Models

The revealed preference travel cost models essentially involve estimating a demand function for recreation. As such, the underlying theory is that of consumer demand theory. A visitor is assumed to maximize their utility subject to a budget constraint. Much like consumer demand theory, there are a number of admissible utility functions which result in different demand specifications. Besides the own price of visiting the recreation site of interest, these demand functions should ideally include the visitor's income and the price of visiting substitute sites. The details of how this conceptual demand model is implemented are specific to the different forms of the travel cost models which we will now be reviewing.

49.5.1 Trip Frequency Models of Recreation Demand

While many public recreation sites have no entrance fee or a minimal administratively set fee, nearly all the implicit price paid for access to the recreation site is the travel cost incurred by the visitor. Thus, travel costs act as a proxy for price in estimating the demand curve. The use of travel cost as a proxy for price hinges on a couple of key assumptions: (a) all travel costs are incurred exclusively to visit this site, and only this site on a trip from home; and (b) there are no significant benefits derived from the travel enroute to the recreation area, i.e., the sightseeing on the way to the site has little value. To meet assumption (a), visitors are queried if they are visiting multiple sites on the same trip and, if so, excluded from the estimation data in most simple trip frequency models but can be included in more complex trip frequency models (Loomis et al. 2000).

Travel cost models employ cross-sectional data that uses spatial variation in visitors' travel costs. There is variation in visitors' travel costs because visitors live at varying distances from the site. With a trip frequency model, the dependent variable is the number of trips each visitor takes over the year or the season to a particular recreation area. The price variable includes the transportation costs (e.g. gasoline), but there may be other variable costs of the trip that would be included in the travel cost variable. These might include lodging or camping fees. Other variables that are usually included as an independent variable in a travel cost model include the visitor's travel time to the site. However, sometimes this variable will be so highly correlated with travel cost that it cannot be included by itself. In that case, the monetary opportunity cost of this time is used to combine the cost of travel time with the transportation cost. Since we are estimating a demand function, other independent variables such as visitor income are usually appropriate to include. Ideally price of the nearest substitute site would be included as well, although this variable is often so correlated with travel cost to the site under study that it is difficult to include. Visitor demographics are also useful as other explanatory variables to act as proxies to control for differences in tastes and preferences.

As single-site trip frequency model is useful if the analyst is interested only in (a) what is the value of current recreation at the site and (b) what would be the loss

in consumer surplus if the site were closed due to agency budget cuts or reallocation of the land to an alternative use (e.g., mine). An example single-site demand curve specification is given in Eq. (49.4) for visitor i:

$$AnTrips_i = \beta_0 - \beta_1 TC_i + \beta_2 TTime_i + \beta_3 Income_i \quad (49.4)$$

where $AnTrips_i$ is annual trips of visitor i to the site, TC is roundtrip travel cost of visitor i, $TTime$ is travel time in hours of visitor i, and $Income$ is household income of visitor i. To address the limitations of this single-site model, a multiple-site model can be estimated. We now turn to a discussion of one such type of multiple-site model.

A multiple-site trip frequency model allows answering a wider range of policy and management questions including how WTP would change for changes in environmental quality or size of the recreation area protected. In order to observe how visitation changes with size of the water body or facilities or environmental quality, there must be variation in recreation site quality or characteristics. While at one recreation site, these attributes are generally fixed, these characteristics usually vary across sites. Therefore, if the analyst pools or combines visitation data from several recreation areas which have varying levels of these attributes, then visitor response to these attributes can be estimated in the demand coefficients. This allows the analyst to estimate how the demand curve shifts with more of a desirable attribute. The area between the original demand curve and the demand curve with increased size or level of environmental quality provides an estimate of the incremental or additional WTP for the increased amount of the attribute. This feature allows the analyst and manager to answer a wide range of policy relevant questions: (a) How the recreation benefits would change with management enhancements such as additional facilities, clean up of water quality, or wildlife management. These marginal benefits can be compared to the marginal costs of carrying out the management action to determine if the added benefits justify the added costs; (b) The change in site quality with allowing an incompatible use to occur at or nearby the site, such as drawing the reservoir level down for irrigation, reducing river flows to produce hydropower, or allowing a nearby mine which would add pollution to a lake. Equation (49.5) specifies what a stylized multiple-site trip frequency demand model would look like for individual i visiting site j:

$$AnTrips_{ij} = \beta_0 - \beta_1 TC_{ij} - \beta_2 TTime_{ij} + \beta_3 Income_i + \beta_4 SS_j + \beta_5 SQ_j \quad (49.5)$$

where $AnTrips_i$ is annual trips of visitor i to the site j, TC_{ij} is round trip travel cost of visitor i to site j, $TTime_{ij}$ is travel time in hours of visitor i to site j, $Income_i$ is household income of visitor i, and SS_j and SQ_j are site size of site j (e.g., number of acres) and site quality of j (e.g., water clarity, fish catch), respectively. The coefficients on the site quality variables indicate how trip changes with a one unit change in site quality. That is, how much the demand curve will shift with a one unit change in site quality? It is from this shift in the demand curve which allows calculation of the marginal benefit of the quality change. This calculation is done

by integrating the area between the current and changed (positively or negatively) demand curve and expanding that to the population of visitors at the site.

There are several econometric specifications of trip frequency models. Historically most trip frequency models were estimated with ordinary least squares regression. However, since 1990, count data regression models have been used since the number of trips taken is a nonnegative integer. Count data models include the Poisson and the Negative Binomial. Negative Binomial count data models do not require that the mean of trips to equal the variance of trips as do the Poisson model.

Since count data models are exponential models, they are equivalent to the semilog of the dependent variable functional form. As such the consumer surplus per trip is simply the reciprocal of the Travel Cost coefficient (Creel and Loomis 1990). See Parsons (2003) or Haab and McConnell (2002) for more details on the count data models.

49.5.2 Multisite Selection Models

Since the 1990s, multiple-site selection models have become popular. These models view the potential visitor as selecting a site to visit from a large choice set of possible recreation sites. These sites differ in terms of travel cost to the site and each sites quality. The individual is assumed to select the site which maximizes their utility given their budget constraint. A repeated discrete choice model has the visitor repeatedly making this site selection decision for each choice occasions (e.g., weekend) over the season and then sums up these trips over all choice occasions in a season.

The theoretical foundation of this model is known as a random utility model since not all the variables in the visitor's utility function are believed to be observable to the analyst. Thus, some of these unobservable variables are treated as random by the analyst, hence the name random utility model. Nonetheless, the site selected by the visitor reflects the one site on any given choice occasion that the visitor views as having the highest net utility. By dividing this utility by the coefficient on travel cost (which is also interpreted as the marginal utility of income), a monetary measure of WTP is calculated. The versatility of this model is that being a multisite model it can value changes in site quality or closure of one or more sites. The strong suite of this model is ability to reflect the influence of substitute sites in the choice of a site to visit. Thus, the loss of value with closing one site is just the incremental loss in utility from having to visit their second best site.

The econometric specification of multisite selection models is quite different from that of trip frequency demand models. Now the dependent variable of the site visited on a particular choice occasion takes on a value 1, and the remainder of sites in the choice set takes a value of zero on that choice occasion. A discrete choice or qualitative response model such as multinomial logit is often estimated. With this model, an increase environmental quality at one site (call it site A) is reflected by some visitors switching away from other sites to visit site A. By linking multinomial

logit site choice model to the trip frequency model discussed above, the analyst can also estimate the benefits of a change in site quality on both site selection and trip frequency. Herriges and Kling (1999) as well as Haab and McConnell (2002) and Parsons (2003) provide an in-depth discussion of these models.

49.5.3 Data Requirements for Travel Cost Models

Obtaining the individual level trip making and travel cost data for travel cost models usually requires a survey of visitors. If a single-site model is being estimated, the task is quite simple since only one site must be visited to collect the data or obtain names/addresses of visitors to send a survey to. However, with multiple-site trip frequency or site selection models, visitation data is needed on many sites. This then increases the data collection costs, especially if on-site surveys are to be used. Alternatively, it may be possible for some activities such as hunting or fishing where licenses are required to do a mail survey and ask the user about all the sites they visit in one survey. This is of course burdensome on the respondent and may reduce the overall survey response rate. However, the payoff from such a detailed survey is the ability to value changes in site quality and account for availability of substitute sites when calculating the demand function and consumer surplus.

49.6 Stated Preference Models

When the change in environmental quality is outside of the prior observed range or the desired value is one of nonuse, then the analyst cannot rely upon actual behavior as there is none. However, economists can construct or simulate a market or a voter referendum to ask people how much they would pay if quality was improved or a unique natural environment protected. The first stated preference method is called contingent valuation method.

We first discuss the contingent valuation method and then a newer stated preference method called the conjoint or choice experiment method. The two stated preference methods share many similarities in that (a) a resource scenario is described to respondents in words, often supplemented by graphs, diagrams, drawings, or pictures to clearly communicate what the resource being valued is and the quantity and quality of that resource. The scenario includes a baseline status quo with no additional cost or no tax cost, and then one or more action alternatives with an associated cost; (b) a means of payment by which the respondent pays the cost of provision of the increased quantity or quality of the natural resource or public good. The means of payment is tailored to the scenario, such that if it is nonuse some form of increased taxes (income, sales, property) or utility bill would be explained as being the mechanism in which the increment of the public good is financed; (c) the WTP question is typically a discrete choice with the respondent being asked if they personally would pay this amount (e.g., in the recreation setting) or vote to pay this amount (e.g., in a public goods setting). The magnitude of the monetary amount

varies across the sample, allowing a quasi-inverse demand curve to be estimated. Given the discrete nature of the WTP question, a logit or probit model is often estimated in order to calculate the maximum amount a respondent would pay.

49.6.1 Contingent Valuation Method

Typically, contingent valuation method is used to estimate a single WTP value for a single scenario offering just one combination of quantity and quality of a public good. For example, in the Exxon Valdez oil spill contingent valuation study (Carson et al. 2003), a one-time WTP for the single scenario of avoiding another equivalently large and damaging oil spill was elicited using in-person interviews. However, some contingent valuation surveys provide multiple scenarios along a common quantity or quality scale. Then, a series of WTP questions are asked, allowing estimation of a WTP function for that increasing quality or quantity of a public good. For example, Walsh et al. (1984) asked annual WTP for four different amounts of land protected as wilderness. Multiple regressions were then used to estimate WTP as a function of acreage protected along with demographics of the visitor.

In terms of the format of the WTP question, Carson et al. used the closed-ended approach in its in-person interviews where respondents were asked if they would pay a particular monetary amount which varied across the sample. Typically at least five, and more often ten, different levels of the monetary amount are asked so as to estimate the quasi-inverse demand curve. An example scenario and a binary closed-ended or *dichotomous choice* referendum WTP question format used by Loomis (1996) for dam removal contingent valuation survey is:

“If an increase in your federal taxes for the next 10 years cost your household \$YY each year to remove the two dams and restore both the river and fish populations would you vote in favor? YES NO”

The \$YY were 15 different bid levels ranging from \$3 to \$190, with most of the bid levels being in between \$15 to \$45.

To estimate the quasi-inverse demand curve, a binary logit model of the following stylized form might be estimated as in Eq. (49.6):

$$\log(\text{Prob Yes}/(1-\text{Prob Yes})) = \beta_0 - \beta_1(\$Bid) + \beta_2X_2 + \beta_3X_3 \quad (49.6)$$

where \$Bid are the \$YY levels asked of the particular respondent, Xs are the values of the non-bid independent variables that may represent tastes and preferences toward the resource of interest.

From this equation, median WTP are calculated following Hanemann (1984) as

$$\text{Median WTP} = (\beta_0 + \beta_2X_{2m} + \beta_3X_{3m})/|\beta_1| \quad (49.7)$$

where X_{2m} , X_{3m} , ..., X_{nm} are the means of the non-bid Xs. Collectively β_0 plus the sum of all the products is sometimes called the grand constant.

Many early contingent valuation method studies from the late 1970s through the 1980s used an open-ended WTP question format where the individual writes into the survey the maximum amount they would pay. This can be analyzed using simple descriptive statistics or ordinary least squares regression. Another popular technique for mail surveys is the payment card, where individuals' circle one of the preprinted monetary amounts representing the maximum amount they would pay. See Boyle (2003) for a more complete description of these alternative WTP question formats and Haab and McConnell (2002) for a detail of the econometric models associated with these question formats.

If use values are obtained, these values are expanded to the user population. For example, if the WTP of asthmatics to reduce air pollution is obtained, the sample would generalize to exogenous estimates of the number of asthmatics in the population of interest. However, if nonuse values such as existence values for a pure public good like protection of the Grand Canyon or an endangered species, the relevant public could conceivably be the entire country.

For the interested reader, a recent but edited book on contingent valuation is Alberini and Kahn (2006). This book provides chapters that included updated guides for designing and implementing a contingent valuation survey, econometric methods, and applications of contingent valuation.

49.6.2 Choice Experiments

In some cases, policy makers do not have a well-defined single scenario but rather are interested in the values of individual natural resource management options that they might combine into an overall management program or project. For example, when restoring wetlands, emphasis could be placed on providing endangered species habitat, but this might require prohibition of all hunting, wildlife viewing, and camping. Alternatively, the area could be managed for waterfowl hunting in one area, wildlife viewing in another, and camping in another part of the wetland. Each of these management options has different direct monetary costs and opportunity costs in terms of other options. Policy makers and managers want to know which of the many possible combinations of management actions would yield the greatest overall net benefits. Choice experiments are designed to answer these questions by estimating the marginal values or part worths of each management option or attribute.

Thinking about this from the viewpoint of the marketing literature, where this method originated, different combinations of management options yield different "product profiles." In our example below, Restoration Option A is 200 acres, with 100 % T&E species habitat and zero hunting and viewing. Restoration Option B might be 200 acres of wetland with one-third available for waterfowl hunting, two-thirds for wildlife viewing, and zero for T&E habitat. The No Action (status quo) or Current Situation usually has a zero cost and serves as a baseline. In our example, the area may currently a "de facto" wetland caused by excess agricultural drainage and used primarily as a "duck club" for private hunting. These product profiles are laid out in choice sets in a table such as [Table 49.1](#)

Table 49.1 Example choice set #1

Allocation of restored wetland	No action/current situation (acres and % of land)	Restoration option A (acres and % of land)	Restoration option B (acres and % of land)
T&E species	0 acres and 0 %	200 acres and 100 %	0 acres and 0 %
Hunting	160 acres and 80 %	0 acres and 0 %	66 acres and 33 %
Viewing	40 acres and 20 %	0 acres and 0 %	134 acres and 67 %
Annual cost per taxpayer	\$0.00	\$50.00	\$75.00
Choose one	[]	[]	[]

Table 49.1's "Choose One" is typical of most Choice Experiments and consistent with the standard random utility formulation that underlies most choice experiments and recreation site selection models. However, this Choose One format does not obtain a great deal of valuation information from each choice, i.e., it is statistically inefficient. One solution typically used is to ask a respondent several of these choice sets. The pros and cons of this approach are briefly discussed below.

In a choice experiment, there are a large number of possible combinations of attributes. This yields a large number of possible choice sets, the exact number depending on how many levels of the four attributes. If there are eight levels of costs to get a precise estimate of the critical "price coefficient" and five levels of the other three attributes, there are dozens of possible combinations in what is called a full factorial design. A more compact design with fewer combinations is a fractional factorial design such as an orthogonal design usually focusing on just the main effects (or what will be the regression coefficient on that variable). In our example, a main effects design has 24 different product profiles, i.e., Restoration Options. The particular 24 combinations that minimize colinearity among the attribute levels are often determined using a SAS statistical software procedure (e.g., OPTEX) or other design choices discussed in Louviere et al. (2000).

The next design decision is how many of these 24 combinations to give each respondent. Generally respondent fatigue begins to set in after answering four such choice sets, and most authors argue against using more than eight (Holmes and Adamowicz 2003).

Once the survey versions are assembled and administered, analyzing the resulting data depends on the format of the choice question. Our example in **Table 49.1** is typical with three options per choice set, so a multinomial logit model is usually estimated when there are three or more options in a choice set. If there are just two options, the current situation and one "action" option, then analysis of this is similar to dichotomous choice contingent valuation and uses a binary logit model.

The multinomial econometric specification for a choice example as depicted in **Table 49.1** would be

$$\text{Prob}(i|3) = \frac{\exp(\beta_0 + \beta_1 A_{i1} + \beta_2 A_{i2} + \beta_3 A_{i3} + \beta_p (\$A_i))}{\sum \exp(\beta_0 + \beta_1 A_{j1} + \beta_2 A_{j2} + \beta_3 A_{j3} + \beta_p (\$A_j))} \quad (49.8)$$

$j = 1, 2, 3$

where in our example with just three choices the sum is of the three alternatives. Essentially the individual is comparing the value of the non-cost attributes with the cost attribute to select the bundle that maximizes the relative utility in option 1 versus options 2 and 3 in our example. See Holmes and Adamowicz (2003) for a more in-depth treatment of the econometric models for these types of choice experiment data.

Once the coefficients from this equation are estimated, the marginal values of each attribute are calculated by dividing the attribute coefficient by the coefficient on cost. With this estimated Eq., the economic values of the different management options can be calculated. Comparing all the values of the different management options allows the analyst to determine the particular combination yields the greatest value. As sometimes happens, the choice experiment survey may have to be conducted prior to managers exogenously arriving at their preferred option based on other criteria. However, once the preferred option is known, the choice experiment results could be used to value that management option for a benefit-cost analysis of that preferred option. This flexibility to value options not identical to what was asked in the survey is also an advantage in many benefit transfers (see Rolfe and Bennett (2006) for a discussion of the advantages of choice experiments for benefit transfer).

49.6.3 The Issue of Bias in Stated Preference Surveys

A commonality of all stated preference methods is the concern about hypothetical bias, i.e., that the stated WTP is not equal to their actual WTP. If hypothetical bias exists, stated WTP is not a valid indicator or “true” WTP. Economists have been concerned about and have studied hypothetical bias for decades. Nonetheless, the issue leaps to the mainstream of economics during the early 1990s when contingent valuation was being applied to estimate the reduction in passive use values from the Exxon Valdez oil spill in Alaska. With hundreds of millions of dollars at stake, the strengths and weaknesses of the contingent valuation method were debated in the *Journal of Economic Perspectives*. Those interested in the debate should see Portney (1994), Hanemann (1994), and Diamond and Hausman (1994).

While the literature on hypothetical bias is voluminous (see Loomis 2011 for a summary), a few key results are worth noting. First, with use values, the bias is not always present. Studies that compare revealed preference techniques such as the hedonic property method to contingent valuation method show no statistical difference in WTP (Brookshire et al. 1982). Comparisons of benefit estimates from travel cost models and the contingent valuation method for recreation use values show, on average, no hypothetical bias (Carson et al. 1996). However, the less familiar the person is with the good being valued the more likely hypothetical bias. Thus, public goods that are largely existence or passive use values for which people do not have firsthand knowledge or prior choice experience do show significant hypothetical bias (Champ et al. 1997).

In response to this hypothetical bias, efforts have been made in survey design to reduce it via exhortations to respondents to behave as if it is a real market where they really have to pay their own money. Ex post calibrations of WTP values

derived from the contingent valuation method have also been proposed based on respondent uncertainty (Champ et al. 1997).

Several other stated preference survey instrument design issues have been labeled as biases. One frequent concern here is payment vehicle bias. This bias occurs if WTP is influenced by how a respondent pays, e.g., via an income tax versus a utility bill. WTP elicitation format bias occurs if WTP is influenced by whether the valuation question is asked in an open-ended format or closed-ended format such as a dichotomous choice or payment card format. See Boyle (2003) for a discussion of these other biases.

49.7 Combining Stated and Revealed Preference Methods and Data

Both stated preference and revealed preference have their strength and weaknesses when estimating use values such as those that might arise from reductions in urban air pollution. Cameron (1992) was one of the first to recognize that perhaps combining revealed preference and stated preference data in environmental valuation might capitalize on their respective strengths while minimizing their weaknesses. In particular, Cameron (1992) talked about using the revealed preference data to “discipline” the stated preference data. This might help reduce the influence that any hypothetical bias might have in the WTP estimates. The marketing literature had been using this approach for more than a decade for a number of purposes including testing for hypothetical bias (see Louviere et al. 2001).

Since the early 1990s, there has been an explosion of combined revealed preference and stated preference studies, particularly in the recreation context. The most recent compendium of state-of-the art papers on combining revealed and stated valuation approaches is Whitehead et al. (2011). This book illustrates the wide variety of applications that the combined revealed preference and stated preference method has been used for. These include pesticide risk reduction, seafood, reservoir operations, as well as recreation.

Of course a reasonable question one might ask is “If you have revealed preference data, why would you want to combine it with stated preference data?” There are several reasons, all related to limitations in relying solely on revealed preference data: (a) revealed preference data may not have sufficient natural variation in amenities or environmental quality to estimate a statistically significant coefficient. This could arise because of limited data availability (e.g., only 1 year of data rather than a time series being available) or because there just isn’t much natural variation in the quality or amenity attribute; (b) the attributes are highly correlated in the data set so that it is nearly impossible to estimate a statistically significant coefficient on each of them separately (e.g., air quality and traffic congestion); (c) the policy being valued would result in changes in quality or level of the amenity that is outside the current range of quality; and (d) introduction of a private good with a new attribute (e.g., locally grown organic corn) or new public good, similar to but not identical with existing public goods.

49.8 Benefit Transfer

Oftentimes economists with state and federal agencies are asked to perform a “quick and dirty” back of the envelope benefit-cost analysis to provide a rough estimate of the benefits and costs of a particular time-sensitive policy proposal or where there is not a sufficient budget that the cost of a survey is feasible.

In this case, environmental economists have developed a set of protocols to transfer existing valuation estimates from prior revealed preference and stated preference studies to evaluate the new policy in question. There are basically four main types of benefit transfer: (a) point estimate value transfer from the most similar study, (b) an average of the values from the prior literatures’ most similar studies, (c) transferring the demand or WTP function from the prior study to the new policy study, and (d) using a *meta-analysis* regression equation estimated on the past valuation studies to calculate what the valuation per unit (e.g., visitor day, household) would be at the new policy site.

In principle, demand/WTP function transfer or meta-analysis have the advantage of being able to adapt the values from the existing literature to better match the criteria for an ideal benefit transfer than would a simple transfer of average values from the literature. Transferring a WTP function that contains demographic variables such as income and age would allow the demand function to be tailored to the sociodemographics surrounding the policy site. In principle, the WTP function approach should reduce benefit-transfer errors as compared to transferring point estimates.

Meta-analysis involves a regression with the value per unit (e.g., recreation day, acre of wetland, household) as the dependent variable and study site characteristics as the independent variables. There have been more than a dozen meta-analyses of environmental and natural resources including water (quality and quantity), electricity, value of statistical life, transportation noise and property values, and wetlands (see Nelson and Kennedy (2009) for a complete listing).

Using a meta-analysis regression equation as a benefit-transfer tool has three potential advantages over average value transfer in terms of an ideal benefit transfer: (a) ability to interpolate a value for a particular public good in a particular region that might not exist in the published literature (e.g., fish species X in region Y might not be available in the literature, only fish species Z in region Y or fish species X in region R); (b) ability to incorporate a nonlinear relationship between the value per unit and the quantity change (e.g., additional acres of wetlands may not have a constant value per acre as an average value transfer implicitly assumes); and (c) ability to account for other attributes of the good being valued (e.g., distinguishing between the value of a recreation activity on public land vs private land). Meta-analyses for benefit transfer are discussed in more detail in Bergstrom and Taylor (2006).

Interest in determining the accuracy of benefit transfer and especially comparing the accuracy of meta-analysis and average value transfers has spawned a substantial literature. This literature uses a comparison of original study values versus benefit-transfer estimates of those same values to calculate the error of benefit transfer. Rosenberger and Loomis (2003) catalog the various estimates of benefit-transfer

errors from value transfers (e.g., point estimates or average values) and function transfers (e.g., primarily meta-analyses). While most of the value transfer errors are in the range of 4–40 %, several are off by 100–200 % (and occasionally more). Benefit function transfer generally does better, but it too can be off by 200 % or more.

One of the tools for improving the accuracy of benefit transfer is for analysts to have access to comprehensive databases and benefit functions. Significant progress has been made in this area in the last two decades. A major advance came with the cooperative Australia, Canada, France, New Zealand, UK, and USA's Environmental Values Reference Inventory (EVRI see <http://www.environment.nsw.gov.au/publications/evri.htm>). This database includes air quality, water quality, wildlife, recreation, and infrastructure. General recreation value databases include Loomis (2005 at http://www.fs.fed.us/pnw/pubs/pnw_gtr658.pdf). For average value tables, databases, and meta-analyses for hunting, fishing, wildlife viewing, wetlands, salmon, endangered species, and open space, see <http://dare.colostate.edu/tools/benefittransfer.aspx> or http://www.defenders.org/programs_and_policy/science_and_economics/conservation_economics/valuation/benefits_toolkit.php.

For the most recent comprehensive discussion of benefit transfer, see the special issue of Ecological Economics edited by Wilson and Hoehn (2006) and Rolfe and Bennett (2006) for a discussion of using Choice Experiments for benefit transfer.

Overall, each of these benefit-transfer methods have their strengths and weaknesses, and the choice is sometimes driven by the (lack) of available data. For example, if there are no similar studies for a similar geographic region, then a meta-analysis may be the best answer if a meta-analysis has already been previously estimated by someone else. If not, then an average of past valuation studies might be the best estimate the analyst can use given the time available to conduct the benefit transfer.

However, any of these benefit-transfer approaches is likely better than omitting completely a monetary value for that health effect or recreation activity. Oftentimes the net result of such an omission from the benefit-cost analysis is an implied value of zero. Benefit transfer, while not as accurate as conducting a primary study, is typically more accurate than an estimate of zero.

49.9 Conclusions

The gist of this chapter can perhaps be summed up in a few sentences. Economic theory provides a consistent measure to value market goods and nonmarket environmental externalities and public goods. Market price is just willingness to pay for one more unit. Where price does not exist, economists can infer willingness to pay using revealed preference methods or using a “constructed or simulated market” ask respondents to state their willingness to pay. The revealed preference and stated preference methods are based on the same utility maximization process economists use to estimate demand for market goods. While the econometric details of estimating an econometric model for recreation are slightly different than estimating the demand for gasoline, often times the basic structure of the data

(e.g., cross-sectional data) and econometric issues dealt with have more in common than one might think.

What cannot be summed up in a few sentences is the wide variety of variations on these basic revealed and stated preference methods. These variations arise due to the need to tailor the valuation to the particular types of public goods. As highlighted in this chapter, the economists' toolkit has a wide variety of methods that can be applied to value nearly every type of public goods that are commonly dealt with in benefit-cost, policy, or regulatory analyses.

Environmental valuation theory and methods are evolving areas of research. While environmental valuation originated in the desire to value recreation in public water projects, it quickly saw application to value air and water quality in benefit-cost analyses of environmental regulation. Environmental valuation rose onto the popular presses radar screen with the application of valuation methods to natural resource damage assessments, including oil spills. In the last decades, as the recognition has grown that the environment provides valuable ecosystem services to people, all the valuation methods discussed above, and stated preference methods in particular, have been employed to monetize these values. Interest in developing computer packages to allow government agencies to monetize ecosystem services relies extensively upon benefit transfer. Environmental valuation techniques continue to see new policy applications and no doubt there will be many more in the future.

References

- Abdullah S, Markandya A, Nunes P (2011) Introduction to economic valuation methods. Chapter 5. In: Batabyal A, Nijkamp P (eds) *Research tools in natural resource and environmental economics*. World Scientific, Hackensack, pp 143–188
- Alberini A, Kahn J (2006) *Handbook on contingent valuation*. Edward Elgar, Northampton
- Anselin L (1988) *Spatial econometrics: methods and models*. Kluwer, Norwell
- Arrow K, Solow R, Portney P, Leamer E, Radner R, Schuman H (1993) Report of the NOAA panel on contingent valuation. *Fed Reg* 58(10):4602–4614
- Bergstrom J, Taylor L (2006) Using meta-analysis for benefits transfer: theory and practice. *Ecol Econ* 60(2):351–360
- Boyle K (2003) Contingent valuation in practice. In: Champ PA, Boyle KJ, Brown TC (eds) *A primer on nonmarket valuation*. Kluwer, Norwell, pp 111–170
- Brookshire D, d'Arge R, Schulze W, Thayer M (1982) Valuing public goods: a comparison of the survey and hedonic approaches. *Am Econ Rev* 72(1):165–177
- Cameron T (1992) Combining contingent valuation and travel cost data for the valuation of nonmarket goods. *Land Econ* 68(4):302–317
- Carson R, Flores N, Martin K, Wright J (1996) Contingent valuation and revealed preference methodologies: comparing the estimates for quasi-public goods. *Land Econ* 72(1):113–128
- Carson R, Mitchell R, Hanemann M, Kopp R, Presser S, Ruud P (2003) Contingent valuation and lost passive value: damages from the Exxon Valdez oil spill. *Environ Res Econ* 25(2):257–286
- Champ P, Brown T, McCollum D (1997) Using donation mechanisms to value nonuse benefits from public goods. *J Environ Econ Manage* 33(1):151–162
- Creel M, Loomis J (1990) Theoretical and empirical advantages of truncated count data estimators for analysis of deer hunting in California. *Am J Agri Econ* 72(2):434–441
- Diamond P, Hausman J (1994) Contingent valuation: is some number better than no number? *J Econ Perspect* 8(4):45–64

- Freeman M (2003) The measurement of environmental and resource values: theory and methods, 2nd edn. Resources for the Future Press, Washington, DC
- Haab T, McConnell K (2002) Valuing environmental and natural resources: the econometrics of non-market valuation. Edward Elgar, Northampton
- Hanemann M (1984) Welfare evaluations in contingent valuation experiments with discrete responses. *Am J Agric Econ* 66(3):332–341
- Hanemann M (1994) Valuing the environment through contingent valuation. *J Econ Perspect* 8(4):19–43
- Herriges J, Kling C (eds) (1999) Valuing recreation and the environment: revealed preference methods in theory and practice. Edward Elgar, Northampton, MA
- Holmes T, Adamowicz W (2003) Attribute-based methods. In: Champ P, Boyle K, Brown T (eds) A primer on nonmarket valuation. Kluwer, Norwell, pp 171–220
- Krutilla J (1967) Conservation reconsidered. *Am Econ Rev* 57(4):777–786
- Loomis J (1996) Measuring the economic benefits of removing dams and restoring the Elwha river: results of a contingent valuation survey. *Water Resource Res* 32(2):441–447
- Loomis J, Yorizane S, Larson D (2000) Testing significance of multi-destination and multi-purpose trip effects in a travel cost method demand model for whale watching trips. *Agr Resource Econ Rev* 29(2):183–191
- Loomis J (2005) Updated outdoor recreation use values on national forests and other public lands. General technical report PNW-GTR-658. Pacific Northwest Research Station, USDA Forest Service, Portland
- Loomis J (2011) What's to know about hypothetical bias in stated preference valuation studies. *J Econ Survey* 25(2):363–370
- Louviere J, Hensher D, Swait J (2001) Stated choice methods: analysis and applications in marketing, transportation and environmental valuation. Cambridge University Press, Cambridge
- Mueller J, Loomis J (2008) Spatial dependence in Hedonic property models: do different corrections result in economically significant differences in estimated implicit prices. *J Agric Res Econ* 33(2):212–231
- Nelson J, Kennedy P (2009) The use (and abuse) of meta-analysis in environmental and natural resource economics: an assessment. *Environ Res Econ* 42(3):345–377
- Parsons G (2003) The travel cost method. In: Champ P, Boyle K, Brown T (eds) A primer on nonmarket valuation. Kluwer, Norwell, pp 269–330
- Portney P (1994) The contingent valuation debate: why economists should care. *J Econ Perspect* 8(4):3–17
- Rolfe J, Bennett J (2006) Choice modelling and the transfer of environmental values. Edward Elgar, Northampton
- Rosenberger R, Loomis J (2003) Benefit Transfer. In: Champ P, Boyle K, Brown T (eds) A primer on nonmarket valuation. Kluwer, Boston, pp 445–482
- Taylor L (2003) The Hedonic method. In: Champ P, Boyle K, Brown T (eds) A primer on nonmarket valuation. Kluwer, Norwell, pp 331–394
- U.S. District Court of Appeals (for the District of Columbia). State of Ohio vs. U.S. Department of Interior (1989) Case number 86–15755. July 14, 1989
- Walsh R, Loomis J, Gillman R (1984) Valuing option, existence and bequest demands for wilderness. *Land Econ* 60(1):14–29
- Whitehead J, Haab T, Huang J-C (2011) Preference data for environmental valuation: combining revealed and stated approaches. Routledge, New York
- Wilson M, Hoehn J (2006) Valuing environmental goods and services using benefit transfer: the state-of-the-art and science. *Ecol Econ* 60(2):335–342

The Hedonic Method for Valuing Environmental Policies and Quality

50

Philip E. Graves

Contents

50.1	Introduction	993
50.2	Value of Statistical Life	994
50.3	Hedonic Valuation of Environmental Quality	996
50.4	Wage Compensation for Environmental Amenities	997
50.5	Property Value Compensation for Environmental Amenities	998
50.6	Wage and Property Value Hedonics Are Not Alternatives: The Multimarket Hedonic Method	1000
50.7	What if Single-Market Hedonic Analyses Are Employed Rather than Multimarket Analyses?	1003
50.8	Conclusions	1007
	References	1008

Abstract

Benefit-cost analysts attempt to compare two states of the world, the status quo and a state in which a policy having benefits and costs is being contemplated. For environmental policies, this comparison is greatly complicated by the difficulty in inferring the values that individuals place on an increment to environmental quality. Unlike ordinary private goods, environmental goods are not directly exchanged in markets with observable prices. In this chapter, the hedonic approach to inferring the benefits of an environmental policy is examined.

P.E. Graves

Department of Economics, University of Colorado, Boulder, CO, USA
e-mail: Philip.graves@colorado.edu; philipegraves@gmail.com

50.1 Introduction

The hedonic approach to valuing environmental benefits has its roots in agricultural economics (see, e.g., Waugh 1928; Vail 1932). Waugh related the price of asparagus, tomatoes, and hothouse cucumbers to various dimensions of perceived quality (e.g., for asparagus, color, size, and uniformity of spears). In another agricultural context, the value of agricultural land has been empirically related to soil fertility and distance from market and more recently to ecosystem services. In yet another agricultural context, this method – not yet known as the “hedonic method” – was employed to isolate “quality” changes from fertilizer price indexes as the former related to changing percentages of nitrogen N, phosphoric acid P, and potash K (see Griliches 1958 for discussion of the early history).

The method first became known as the “hedonic method” as a result of Andrew Court’s (1939) work at General Motors. Court was interested in separating quality improvements from price increases as automobiles improved rapidly during the early decades after their first introduction. One can implicitly value the horsepower, size, and various other model features with this method, and that valuation could in turn be used to increase GM profit by providing more high-value but low-cost features.

In a now-classic article, Solow used what was essentially the hedonic method in a time series context, holding constant measurable inputs to explain GDP growth – his now-famous “residual” (technological change) was seen to account for a quite large percentage of economic growth, the forerunner of modern endogenous growth models.

It is Griliches (1961), however, who is generally viewed as the “modern father” of the hedonic method. He introduced many refinements in the method in the context of separating quality improvements from price increases to allow construction of better price indices to more accurately measure GDP growth. Early studies tended to focus on either the demand side or the supply side, with Rosen (1974) being the first to present a full general equilibrium discussion; the now-classic Roback (1982) contribution brought the realization that a full general equilibrium requires joint consideration of property value and wage differentials, which we shall return to later in this chapter.

The earliest environmental application of the hedonic method was that of Ridker and Henning (1967). They established that housing prices in St. Louis were higher in cleaner areas, other things equal. There has been a proliferation of property value studies since that time. Valuing water quality is somewhat more difficult with the hedonic method for reasons beyond the scope of this chapter, and far fewer studies have been conducted for this environmental media. A relatively limited number of studies have also attempted to value noise from highways and airports as well as hazardous waste dumps. The valuation of each type of environmental amenity generally brings an amenity-specific set of problems, although the focus here will be primarily on air pollution.

50.2 Value of Statistical Life

This section discusses the first of two distinct areas in which the hedonic method is used in environmental policy, while the section to follow deals with the second. The “value of statistical life” (VSL) is useful to value mortality damages (in the “health effects” or “sum of specific damages” approach) in public policies of a wide variety, environmental policies being emphasized here. This method employs wage regressions to value the risk of on-the-job death, with more risky jobs requiring higher wages, at least in principle. In these studies, the dependent variable is wages (or \ln wages) of individual workers which is regressed upon a vector of individual personal characteristics (e.g., age, education, race, sex, experience) and job characteristics (e.g., occupation, industry, unionization). The risk of death, although quite controversially measured, is then included, with an expectation of a positive coefficient to reflect the needed compensation for job risk.

The compensation required for the higher risk can then be used to estimate the VSL for use in broader policy contexts. Suppose, for example, there is a 1/100,000th higher annual probability of dying on the job as a lumberjack than in an average job and that the typical lumberjack (of, say, 100,000 total) required \$50 more per year (2.5 cents/h, with a 2,000 h-year) to accept this risk. The expected number of excess deaths is then one, and the aggregate willingness-to-pay (the VSL) would be \$5 million, a number not far from those used in actual public policies. If a particular policy is expected to save 20 lives, with no other effects, it would have \$100 million in benefits to be compared to provision costs.

VSL has been inferred in non-labor market settings, as well, with the purchase of smoke detectors, seat belt use, various automotive safety features, etc. having been studied. These other approaches are, however, typically undertaken to corroborate the more ubiquitous labor market approach.

Focusing on the wage hedonic work that has been the dominant influence on environmental policy, there are numerous problems with the conduct and interpretation of these studies (see Dockins, et al. 2004 for an excellent, and very complete, review of existing VSL studies and their limitations):

1. Do people perceive low-probability risks at all accurately? Are actuarial risks more or less appropriate to use than perceived risks when the two differ? Are the actuarial risks themselves properly measured (e.g., a common observation is that the National Institute of Occupational Safety and Health (NIOSH) data on risks yields VSL estimates that are substantially higher than those obtained with Bureau of Labor Statistics (BLS) data, and it is likely that there is also substantial risk measurement error *within* each of these basic data sources)?
2. Is the “marginal” worker in a risky occupation more concerned about risk than the “average” worker? If so, as is likely, the VSL will be biased upward by using the marginal worker’s required compensation.
3. Does the functional specification matter (e.g., linear, \ln -linear, Box-Cox, squared terms)? There is little or no theoretical evidence on which functional form is appropriate to apply.

4. Does inclusion or exclusion of other variables affecting wages result in big apparent changes in VSL? For example, risk of non-death injury is likely to be highly correlated with risk of death; omitting the former will bias the latter upward. Black et al. (2003) find the coefficients from the wage hedonic to be highly unstable with respect to both functional form and data selection.
5. Finally, has the Environmental Protection Agency (EPA) ignored potentially important additional concerns? The EPA does not support adjustments to VSL based on how one dies in specific jobs, age, cross-sectional income, non-death risk dread (e.g., cancer), baseline health status, or voluntariness/controllability of risk – yet each of these might be relevant for an individual’s willingness-to-pay for risk reduction. Trudy Cameron (2010) offers a recent balanced view on the nature of VSL which, among other contributions, suggests that a “less incendiary” terminology than value of statistical life be substituted, perhaps “willingness-to-swap (WTS) alternative goods and services for a micro-risk reduction in the chance of sudden death.”

Progress in the estimation of VSL is ongoing, and many of the concerns raised here are being examined in an effort to improve existing VSL estimates, as seen in the Cameron paper. However, a central insight that cannot be escaped is that any policy decision that involves changes in the probability of death *inevitably* represents an implicit valuation on a statistical life. Explicitly using a specific VSL number is quite likely to lead to better decisions and to decisions that can be analyzed to determine how sensitive the benefit numbers are to alternative assumptions about the magnitude of VSL.

50.3 Hedonic Valuation of Environmental Quality

There have been many studies using either (or in rare cases both) property value hedonic equations or wage hedonic equations to value environmental quality. In either approach, the dependent variable (either wages or property value) is regressed upon as many causative independent variables as are reasonably available, to which are added variables measuring environmental quality. Numerous readily available review articles have dealt with the many theoretical and econometric issues with the hedonic method (e.g., Palmquist 2005; Taylor 2008), while Graves (2011) presents a simplified verbal and graphical exposition that is accessible to those with widely varying backgrounds. The approach taken here is to provide a “middle-ground” verbal approach to understanding the hedonic method, an approach that will be seen to clarify some interpretations that are either not widely known or which are ignored in typical studies.

As was the case with VSL studies, accurate valuations of environmental improvements in either land or labor markets require that households have “good” (perfect ideally) perceptions of both (i) where it is clean and dirty and (ii) what various levels of environmental quality mean to our health and welfare. Under such strong assumptions, one would expect people to ponder how to avoid risks of death, on the one hand, or pollution damages, on the other. The insight that

underlies the hedonic approach to environmental valuation is that as long as an individual's marginal cost of avoiding damages is less than the marginal benefits of avoiding damages, that individual would be expected to continue to avoid damages until marginal costs and benefits are equated.

Households can lower pollution damages by either moving to a cleaner town or by moving to a cleaner part of the town they currently occupy. However, since many other movers and non-movers would – other things equal – prefer to occupy cleaner locations, other things cannot remain equal. As will soon become clear, one would expect to observe falling wages and rising housing prices in the clean location until identical households are no better off in a clean location than in a dirty one.

While this central idea is straightforward, there is confusion in the details, a confusion this chapter is designed to clarify. We shall take up the labor market approach in Sect. 50.4, since it follows naturally from the VSL discussion, turning to the property value approach in Sect. 50.5. The only difference between our earlier discussions is that rather than focusing on wage compensation for risks of death on the job, we focus on environmental quality which varies among labor markets, hence should lead to varying levels of wage compensation among those labor markets.

50.4 Wage Compensation for Environmental Amenities

If City A, one of two otherwise equivalent cities, has higher pollution levels than City B, one would expect residents to move from A to B, reducing the labor supply (raising wages) in A and increasing labor supply (lowering wages) in B – and one would expect this movement to continue until the relatively lower wage in B exactly compensates for the utility value of B's better environment (as we shall see later, this expectation is not fully correct).

One powerful advantage of this approach is that the benefits of environmental cleanup are directly observed in dollar terms, which makes for very convenient comparison to the dollar costs of policies that would result in cleaner cities. Moreover, nonlinearities and synergistic interactions among various pollutant types can readily be explored. This can be easily seen with reference to the following estimation equation:

$$W = \alpha + \beta X + \gamma PM_{10} + \theta (PM_{10})^2 + \delta SO_2 + \lambda (SO_2)^2 + \eta (PM_{10}SO_2) + \varepsilon \quad (50.1)$$

where W is annualized (or hourly) wages, X is a complete vector of traditional wage determinants employed in earning functions in the labor economics literature (education, experience, age, occupation, region, union, etc.), and β is the vector of coefficients on the variables in X . PM_{10} is particulate matter 10 μm in diameter or smaller, SO_2 is sulfur dioxide, and the Greek letters preceding these variables are their respective regression coefficients. The error term, ε , of the regression must

meet certain classical regression requirements (iid, no spatial autocorrelation, etc.) with failure to meet those requirements suggesting mis-specification of the regression model. Once a data set, hopefully with many observations, has been amassed and the regression in Eq. (50.1) has been properly specified and estimated, it is a simple matter to calculate marginal pollution damages:

$$\partial W / \partial PM_{10} = \gamma + 2\theta(PM_{10}) + \eta(SO_2) \quad (50.2)$$

The interpretation of Eq. (50.2) is quite simple: the first term, γ , is the marginal damage from an incremental change in PM_{10} under linearity (expected to be positive as discussed earlier); the second term indicates the degree of nonlinearity (e.g., marginal damages are increasing in pollution levels if $\theta > 0$), while the final term indicates the extent to which PM_{10} damages depend on how much SO_2 is present. All of the coefficients would be in convenient dollar form, and to the extent that the second two terms are significantly different from zero, public policy should have pollution standards for any particular pollutant (and economic incentives) that vary with both (a) levels of pollution and (b) levels of other pollutants present. At present, this possibility is completely ignored in environmental policy, and a fruitful line of research would be to delve more deeply into nonlinear and synergistic damages. Since there is very little theoretical guidance on the nature of the appropriate functional form for pollution damages, researchers (inadvertently) and advocates (deliberately) might well distort environmental values by their choices along a number of dimensions (omitting variables that are positively or negatively correlated with the environmental variables, employing a linear model when the data suggest a nonlinear form is more appropriate, etc.).

In closing discussion of the wage hedonic approach, it should be reemphasized that this method only works well when people are very aware of both where it is clean and dirty and how working in a clean or dirty location affects them. Bockstael and McConnell (2007), however, in a review of wage studies, find clear evidence that households are willing to give up wages to live in cleaner locations.

50.5 Property Value Compensation for Environmental Amenities

The property value or rent compensation method employs a virtually identical way of thinking but applies the notion that movements will equilibrate utility within an urban area through adjustments in land values. How much a house will rent or sell for is clearly related to the bundle of positive and negative traits that comprise it. The traits are many: structural (e.g., stone or wood, square footage, number of bathrooms, lot size, type of heat), neighborhood (e.g., school quality, crime rates, access to a wide variety of destinations, notably the central business district in traditional urban models), and – our interest here – environmental quality.

Environmental quality is sometimes viewed as a “public good” in the sense that whatever environment exists in an area is essentially unaffected by an individual

household's behavior and that an individual household cannot be excluded from enjoying whatever level of environmental quality exists in that area. The property value hedonic method relies on the location specificity of pollution levels – that they vary over space in an urban area – to convert environmental quality into a private good that is “bundled” with housing choice. As with the wage hedonic, assuming that perceptions are “good” (ideally perfect), the value of varying levels of pollution within a city should be captured in property values.

The process is quite similar to the wage hedonic approach and can be represented as in Eq. (50.3):

$$PV = \alpha + \beta X + \gamma PM_{10} + \theta (PM_{10})^2 + \delta SO_2 + \lambda (SO_2)^2 + \eta (PM_{10}SO_2) + \varepsilon \quad (50.3)$$

where PV, property value, is ideally actual sale price rather than listing price, with the only important difference from Eq. (50.1) being that instead of containing variables affecting wage, the X vector instead is comprised of all structural and neighborhood traits affecting housing value, with the other variables are as defined earlier. The Greek coefficients are the regression coefficients of a properly specified model resulting in an error term with appropriate properties.

To find how property values vary in a systematic, functional way with pollution levels, we again partially differentiate Eq. (50.3) with respect to a pollutant of interest, say particulates:

$$\partial PV / \partial PM_{10} = \gamma + 2\theta (PM_{10}) + \eta (SO_2) \quad (50.4)$$

The interpretation of the coefficients are exactly as before, with γ capturing the linear impact of pollution on property value, 2θ capturing the extent of any nonlinearities, and η testing for synergisms ($\eta > 0$ damages are “supra-additive,” while $\eta < 0$ damages are “sub-additive”). Krumm and Graves (1982) found a significant positive η indicating synergistic increases in particulate damage, measured by hospital admissions, when more sulfur dioxide is present. As with the wage hedonic, the coefficients give us marginal damages (the benefits of cleaning up) in a very convenient dollar form enabling comparison to marginal provision costs.

As with the wage hedonic approach, there is little theoretical guidance as to the nature of the functional relationship between property values or rents and the traits that exert a causative influence, allowing advocates to intentionally publish widely varying results even from identical raw data. Krumm and Graves employed a methodology devised by Zellner and Siow (1980) that, at least in principle, eliminates biases when theoretical guidance on functional form is limited. The potential to deliberately publish biased results is of more than academic concern since there is considerable evidence that estimated property value effects of pollution are not robust to alternative specifications (see Graves et al. (1988) for more in-depth discussion).

For either of the wage or property value methods, problems related to either data limitations or assumption of perfect information exist. If, for example, some other

disamenities are positively correlated with the pollution measure and those other disamenities are omitted from the equation, the value of the pollution damages will be biased upward. For example, suppose that the more polluted parts of a city are also less desirable for other reasons (more crime, worse schools, more graffiti, street potholes, poorer lighting, fewer parks, etc.) and these other traits are omitted from the equation. By not including the other goods that are correlated with pollution, the impact of pollution will appear to be larger than it is, since the effects of the other non-included variables will be partially attributed to environmental quality (the magnitude of the bias will equal the coefficient on the omitted variable if it were included times the correlation coefficient between that variable and pollution). With constantly improving data acquisition, this problem is becoming less important over time.

Since experts argue heatedly about health and other damages and since many pollutants are odorless, colorless, and tasteless in ambient concentrations, it is plausible that households might fail to fully perceive either (a) the impact of pollution on their health and well-being, (b) how pollution levels vary over space, or (c) both. To the extent that perceptions are imperfect, one would expect that the hedonic methods would yield pollution damage coefficients that are biased downward, since households would not be expected to be willing to pay for unperceived benefits of cleaner locations.

What is the net effect of these potential biases, one suggesting overvaluation and one suggesting undervaluation? Nobody knows the answer to this question with confidence, but a great many property value studies – as was the case with the smaller number of wage differential studies – show strong positive relationship between property values and environmental quality. The property value approach might be thought to be particularly useful for valuing spatially concentrated environmental damages (e.g., toxic waste dumps), and the wage differential approach might seem more appropriate for region-wide amenities (e.g., large pollution clouds or climate). As we shall see in the following section, these beliefs are, generally, quite flawed.

50.6 Wage and Property Value Hedonics Are Not Alternatives: The Multimarket Hedonic Method

Until fairly recently (new information spreads slowly), the two approaches to valuing pollution damages were viewed as alternative approaches. It was thought that clean air, for example, could be valued either by variation in property values within an urban area or by wage variation between urban areas. Indeed, if the values happened to be similar under the two methods, greater confidence was placed in either as a measure.

It turns out that this is incorrect under plausible assumptions about people's behavior when evaluating locations. Indeed, for this view to be valid, households would have to follow a two-stage procedure when locating – first, looking only at wages, select a labor market, and at a second stage, select a location within that

labor market. This would clearly be irrational since households could make better location decisions by looking at the combination of wages, rents, and amenities available in all locations prior to selecting their best location.

To further clarify, another way to think about this is that, between two otherwise identical locations, the one that is more polluted will be less attractive – so people will move from the more-polluted to the less-polluted location until they are equally well off in both locations. But, as they move into the less-polluted location, they both increase the supply of labor (driving down wages) and increase the demand for housing (driving up property values and rents). Hence, the true value of the less-polluted locations is the sum of what must be paid for reduced pollution in both the labor and land markets.

To many, the argument of the preceding paragraph is not clear or convincing, so additional discussion is useful. Suppose, as a soon to be dropped initial assumption, that the entire world were a flat, featureless plain where all locations are identical. In this scenario, there is no variation in closeness to ocean, scenic views, and the like. Just as there would be no reason to pay more for identical automobiles, there would be no reason to pay more in either land or labor markets for one location over another. Further, as again soon to be dropped, assume that all households have the same preferences and all firms have the same cost functions (and are selling on national markets at one price, hence have the same profit functions in all locations). With these assumptions in place, there would be no variation in demand for lots of different sizes or for hours worked on the part of households nor would there be variation in the relative land/labor intensity on the part of firms.

In this simple initial scenario, wages and rents would – in equilibrium – have to be the same in all locations. If, for example, there were a location with higher rents, households would have to be compensated by higher wages or they could not be equally well off there vis-à-vis elsewhere. But, if they *are* compensated with higher wages, the higher wage/rent location would have to be less profitable than other locations for firms; hence, firms would leave, reducing the demand for labor and indirectly reducing land demand as household employment falls. If a location had lower rents than elsewhere, households would move in until lower equilibrium wages made them indifferent to other locations – but the lower rents and wages would stimulate firm in-migration, until wages and rents were raised to those of other locations. Hence, were the world as boring as the flat, featureless plain and homogeneous household/firm assumptions imply, rents and wages would have to be identical in all locations in equilibrium (see Graves 2011 for a full graphical presentation of this and subsequent discussion).

Now let us begin dropping these unrealistic assumptions, first by introducing variation in an amenity that households care about (e.g., a scenic view or lower humidity), but which has no impact on firm profitability. If we are at an initial equilibrium with wage and rent levels equal in all locations, any location possessing more than average amounts of the desirable amenity will be more attractive, hence will lure in-migration of households. But that in-migration will result in increased labor supply along with increased land demand. Hence, the desirable location(s) will

have lower wages and higher rents, in some combination that renders – in equilibrium – utility the same in the desirable locations(s) as in the average locations.

Similarly, undesirable locations will experience household out-migration at the initial common wage and rent levels, resulting in some combination of lower rents and higher wages in undesirable locations relative to average locations. Note that the compensation paid (for desirable locations) or received (at undesirable locations) represents a measure of “quality of life.” The higher rents and lower wages do not represent a higher “cost of living” in the nice locations but rather a higher “benefit of living” there. The higher benefits of living in the desirable location – as with quality variation among ordinary goods – must be paid for in equilibrium. Hence, were all households homogeneous, there would be in equilibrium no reason to prefer one location over another, despite wide variation in amenity levels, since any gain in amenities would be fully offset by higher rents and lower wages and conversely. Locations that are unusually nice for households will have larger populations than other places.

If an amenity affects firm profitability (e.g., access to resource inputs) without having any impact on household utility that will not, in equilibrium, result in greater profits for the firm. Rather, firms will enter driving up land rents directly and indirectly via employment and driving up wages (the latter necessary to compensate households for the higher rents, which is made necessary by the fact that the location is no “nicer” for them). Note that in this case, the higher rents *do* represent a higher “cost of living” but that higher cost must be completely offset by higher wages. Locations that are unusually nice for firms will, as with household amenities, have larger populations than other places.

The preceding two cases lead to nine spatial combinations, with a rich tapestry of possible wage and rent combinations:

- a. The “average” location (average wage, W_0 , average rent, R_0 , and average size, S_0)
- b. Nice for households, neutral for firms (lower W , higher R , larger S)
- c. Bad for households, neutral for firms (higher W , lower R , smaller S)
- d. Nice for firms, neutral for households (higher W , higher R , larger S)
- e. Bad for firms, neutral for households (lower W , lower R , smaller S)
- f. Nice for both households and firms (ambiguous W , higher R , larger S)
- g. Bad for both households and firms (ambiguous W , lower R , smaller S)
- h. Bad for households, good for firms (higher W , ambiguous R , ambiguous S)
- i. Bad for firms, good for households (lower W , ambiguous R , ambiguous S)

Case 9, which we will return to in the following subsection, is of particular interest for environmental policy, since many environmental policies raise the costs of firms but provide benefits to households. Until the early 1980s, most economists believed that imposing stringent controls on firms in a location would result in them leaving that location. This led to fears of a “race to the bottom,” since firms leaving raise unemployment in the short run and firms entering less stringently regulated areas would reduce unemployment in the short run. This presumption was based on a focus on the firm impact, ignoring the impact on

households. If the cost increases associated with the environmental policy are relatively small and the household benefits relatively large, the location might well experience growth as it moves to a larger equilibrium size, S .

Ignored in, but implicit in, the hedonic discussion to this point is the impact of in-migration and out-migration on what might be called “endogenous amenities and disamenities.” That is, if a desirable location for households exists, one would expect in-migration until the lower wages and higher rents rendered that location no more desirable than other locations. But, it is also the case that in-migration might increase levels of endogenous disamenities (e.g., pollution, congestion) or might increase levels of endogenous amenities (e.g., restaurant diversity, local goods with scale economies in production). In a full general equilibrium analysis with all important amenities included, this would not matter because the “net niceness” of the location will still be captured by rents and wages.

But data limitations in actual studies are likely to lead to mismeasurement of the value of amenities. If, for example, measures of increased cultural opportunities or restaurant quality and diversity are positively correlated with the amenity but are omitted from the hedonic estimates, then the value of the amenity will be *overstated* by the wage/rent differentials observed – the coefficient on the amenity variable will be larger by the omitted variables’ effect times the correlation with the amenity. Similarly, if increased congestion is positively correlated with the amenity but was omitted from the equation, then the wage/rent variation would *understate* the value of the amenity. In the environmental context, increases in pollution and congestion are both likely consequences of movements to desirable – perhaps because of better climate – locations, but if changes in congestion are omitted from the estimating equation, the pollution variable will pick up the effect of congestion times its correlation with pollution.

A common criticism of the underlying assumptions of the hedonic method is that households and firms, especially the latter, might have very high movement costs; hence, disequilibrium might persist for very long periods. If this is the case, then observed wage and rent differentials would not be entirely compensatory. That is, high-wage places might be “high-utility” places because more of all goods could be consumed there, while high-rent places might be “low-utility” places because fewer goods could be consumed in such high-cost locations.

A couple of observations are pertinent to this issue. First, it might not take too many people or firms actually moving to yield a close approximation to a full-mobility equilibrium. This is analogous to the fact that only a few drivers need to move from “slow lanes” to “fast lanes” on a freeway at rush hour to make all lanes equally fast. Second, as an empirical matter, in recent decades, households have been moving *toward* high-rent locations and *toward* low-wage locations. With rising nationwide incomes, this trend is consistent with an equilibrium in which desirable locations are also normal or superior goods (i.e., at higher national incomes, there is even greater demand for the already desirable subnational locations).

50.7 What if Single-Market Hedonic Analyses Are Employed Rather than Multimarket Analyses?

Very few multimarket hedonic analyses have been conducted (for an early contribution, see Blomquist et al. 1988), while a very large number of hedonic analyses have been conducted in either the labor or land markets considered separately. What are the implications for environmental valuation of using a rent hedonic or a wage hedonic rather than the combined analysis implied by prior discussion? The taxonomy of household/firm amenity combinations, 1 through 9 above, has clear implications for valuation biases introduced by failure to consider both markets.

We will focus on the policy-relevant case where one is attempting to determine the value of environmental quality to households (to use that information to infer benefits to be compared to costs in environmental benefit-cost analysis). Many environmental policies tend to be applied uniformly over space, but that does not, in general, mean that their benefits and costs are uniformly distributed over space. For example, required catalytic converters on automobiles raised costs in rough proportion to population, but the benefits of that policy would be much higher in locations (e.g., Los Angeles, Phoenix, Denver) that are sunny and warm and/or have stagnant air conditions. Hence, a uniform policy can have pronounced effects in making locations such as Los Angeles *relatively* more attractive than they would otherwise be, encouraging in-migration and resulting higher land values and lower wage rates. In cases such as this, where there are negligible impacts on local firms, it would clearly be the case that using either a property value hedonic or a wage differential hedonic in isolation would result in *underestimation* of environmental benefits relative to a multimarket approach. The extent of bias will depend on the relative capitalization rates and which of the two hedonic methods are chosen – if most of the impact of the policy goes into wages, using a property value approach will greatly underestimate benefits, and conversely.

Other more location-specific environmental policies, such as Pittsburgh introducing controls on steel polluters in the 1950s prior to nationwide control policies, will have direct impacts on both local firms (harmful) and local households (favorable). The harm to firms would lead to lower demand for labor, while the desirable impacts on households would lead to an increase in the supply of labor. Both of these effects cause wages to fall, but the net effect on property values/rents is ambiguous, depending on whether the city gets larger or smaller as a result of the policy. In such cases, a wage hedonic is much more likely to accurately value the environmental improvement than would a rent hedonic, the latter falsely implying little or no environmental benefits from the policy. As a “fluke,” it might be that the wage hedonic picks up the full value of the environmental policy, but in general, adding the information from a property value study would lead to more accurate estimates. It should be noted that the compensation shares are not limited to [0,1], but rather more than 100 % of the benefits could go into wages and the rent compensation could actually be negative (e.g., if the environmental policy harmed firms very much, so that the city got smaller, with lower rents in equilibrium).

If a property value study is used in this case, it would seem like the environmental improvement had negative value!

If an environmental policy at a location is good for both households and firms, both would want to move in. Suppose, for example, that a nationwide law is passed that subsidizes firms to clean up in areas where there is non-attainment of air pollution standards, with no subsidy in areas meeting current air pollution standards. This situation would cause rents to rise in locations subject to the policy with an ambiguous impact on wages. Wages would rise if the policy benefited firms relatively more than households, while wages would fall if the policy benefited households relatively more than firms. In this case, the value of the environmental amenity would appear largely in land markets, but again only as a fluke would there be no labor market effects. In this case, the choice of a hedonic wage analysis will greatly underestimate the value of the cleaner air.

It has been assumed to this point that all households and all firms are homogeneous. This is of course not the case in realistic settings. Land-intensive firms would not be expected to be found in locations where land is expensive (which is why corn is not seen growing in downtown New York City). Similarly, those households who have unusually large preferences for land, perhaps those with large families or pronounced gardening desires, would not locate where land is very expensive, perhaps locating in suburbs or exurbia rather than in more central areas.

If a firm's labor demands are unusually large, it would avoid locations with unusually high wages. If a household does not supply labor (e.g., the retired as discussed in Graves and Waldman 1991), they would want to locate where amenities are mostly paid for in wages rather than rents. This would also be the case for those who have very high demand for services. Conversely, those households that supply low-skilled labor to service industries are likely to be priced out of very desirable and high-rent locations (e.g., Malibu, CA; Aspen, CO; or Key West, FL) and will have to be compensated in higher wages to locate there or commute in to work, that is, the low-skilled may actually have higher wages in desirable locations. As the preceding discussion makes clear and as even casual reference to the real world verifies, there is a very rich tapestry of locational choices when the full implications of the role of firm and household amenities are considered. This is even more the case when endogenous amenities are considered, amenities such as the amount of similar people present in a community (e.g., the ethnic neighborhoods of large cities that often make them much more attractive to particular types of people than would otherwise be the case). Summarizing, there are five reasons why hedonic methods are likely to underestimate the value of environmental quality improvements. The first, and most obviously damaging, is that the benefits of environmental quality must be fully perceived by households for them to be willing to pay more for cleaner locations. As mentioned earlier, even the world's foremost health experts have spirited debates about the role various pollutants play in human disease and death. It seems very implausible that ordinary people would be able to accurately perceive such things. Additionally, many pollutants are odorless, colorless, and tasteless in normal ambient concentrations; hence, ordinary people might

be unable to distinguish the clean places from others. It is unlikely then that many important environmental effects would be capitalized into property values.

Why do hedonic studies show such large environmental effects then? It is certainly the case that people will perceive localized smells, bad visibility, and other impacts of pollution that are inevitably revealed by our five senses. Yet, it is precisely such perceived damages that are ignored in the sum of specific damages approach (sometimes called the “health effects” or “averting behavior” approach) which is often used in environmental policy analysis. A good argument – the second reason why hedonic methods underestimate environmental values – could be made for *adding* the damages estimated via sum of specific damages (lives saved, reduced asthma attacks, etc.) to those estimated via hedonic methods. This follows from the fact that the damage categories measured by the two methods exhibit very little overlap – damages that are *perceived* would be expected to go into property values and wage differentials, while damages that are *unperceived* would be measured by the sum of specific damages approach.

The third reason – discussed at length earlier – for why hedonic methods are likely to underestimate environmental values is that it is still the case that separate analyses in labor or land markets are still the norm, when it has been known for several decades now that only a multimarket hedonic can accurately capture the full value of the environment. The circumstances under which a single-market analysis could accurately value an environmental amenity are extremely rare (e.g., a fixed housing stock, a retired population).

The fourth reason for expecting the hedonic method to underestimate true benefits is that the hedonic method, even properly conducted, only captures *use* benefits of the environmental resources of concern, since the amenities are bundled with housing and jobs. Nonuse benefits might well be of greater magnitude in particular environmental settings, and policies allocating the environmental resource should, on efficiency grounds, encourage highest value usage even if that results in nonuse of the environmental resources. Illustrating, is the California Coastal Commission properly allocating scarce ocean locations? It is clear that in the absence of this regulatory authority, virtually the entire coast of California would be lined with high-rise condos, looking much more like Miami than at present. But, the scenic Pacific Coast Highway has value to all who drive it, and to a large extent, that value has been perceived as being of greater importance than the (admittedly very large) benefits households would receive if the coast were opened to unrestricted development.

The final reason why hedonic methods might be expected to underestimate the benefits of environmental cleanup stems from the supplies of clean locations relative to the demands for clean locations. The hedonic method results, at least in principle, in zero spatial consumer surplus for similar households. That is, if one location is nicer than another location, households will continue to move to the nicer location, until it is no longer nicer, until identical locations have identical prices. There will be no consumer surplus over space, and indeed, this is one of the reasons the hedonic method is desirable in that the full benefits that are perceived are measured.

But the fact that people are very different means that understatement of environmental benefits (damage reduction) can occur if there are more locations with the amenity than there are people strongly desiring the amenity. Suppose, for example, that there are very few households containing really unhealthy individuals, individuals with weakened cardiopulmonary systems who would be highly damaged by pollution. Such households might be willing to pay a great deal for a very clean location, but they might only have to pay a much smaller amount, if the number of somewhat clean locations is large relative to the number of these households. They will get, in other words, consumer's surplus over space. Inferring the value of cleaning up the environment from the average person in this case would ignore the high marginal benefits received by these households. As another illustration of the potential importance of this point, a hedonic analysis of a large city might suggest that its mass transit system has low value, because those who have the greatest use value (e.g., the disabled or those who particularly dislike automobile commuting) may only have to pay a small portion of their true willingness-to-pay in land or labor markets.

When one considers the very large number of traits that can matter to a heterogeneous population with very diverse preferences, it becomes clear that a great deal of consumer surplus can remain in the hedonic equilibrium. In the case of incrementable environmental goods, the unobserved consumer surplus corresponds to a higher marginal value that might if observed justify a policy intervention to increase levels of the public good.

The hedonic method is quite popular due to its ability to provide a convenient dollar measure of marginal environmental damages (damage reduction being the benefit of environmental cleanup policies). The limitations discussed here imply that there is a great deal of room for improvement in this method and raise issues of how best to get at the *total* marginal benefits, measured in all markets that households have to pay in.

50.8 Conclusions

The goal of this chapter was to describe the hedonic method as a means of valuing environmental quality improvements. The hedonic approach requires very good, ideally perfect, perceptions of environmental benefits (or risk in the VSL case) along with good/perfect knowledge of how environmental quality varies over space (or risk over jobs in the VSL case). This assumption is highly suspect in many settings.

Moreover, it remains the case that expert legal testimony and typical regulatory practice still commonly employ either a property value study or a wage study, despite our having known for more than two decades that compensation for environmental amenities and disamenities will generally occur in both the land and labor markets. The extent to which damages appear in land versus labor markets would generally vary according to many things, but considering either market separately is likely to greatly underestimate the damages from pollution. If an environmental pollutant were highly concentrated (e.g., a hazardous waste dump),

one would expect a greater percentage of its damage to appear in property values, while the damages from more regionally ubiquitous pollutants might be expected to appear primarily in wage rates. The existence of firm amenities and disamenities complicates the ability to establish general conclusions, however, but it remains the case that using only one of the two markets in which environmental quality is valued generally results in understatement of environmental values.

References

- Black DA, Galdo J, Liu L (2003) How robust are hedonic wage estimates of the price of risk? Final report to the USEPA (R 829-43-001)
- Blomquist GC, Berger G, Hoehn J (1988) New estimates of the quality of life in urban areas. *Am Econ Rev* 78(1):89–107
- Bockstaal NE, McConnell KE (2007) Hedonic wage analysis. In: Environmental and resource valuation with revealed preferences: a theoretical guide to empirical models. the economics of non-market goods and resources, vol 7, Springer, New York, pp 151–187
- Cameron T (2010) Euthanizing the value of a statistical life. *Rev Environ Econ Policy* 4(2):161–178
- Court AT (1939) Hedonic price indexes with automotive examples. In: The dynamics of automobile demand. General Motors Corporation, New York, pp 99–117
- Dockins C, Maguire K, Simon N, Sullivan M (2004) Value of statistical life analysis and environmental policy: a white paper, U.S. Environmental Protection Agency, National Center for Environmental Economics, April 21. For presentation to Science Advisory Board, Environmental Economics Advisory Committee
- Graves PE (2011) The hedonic method: value of statistical life, wage compensation, and property value compensation. In: Batabyal A, Nijkamp P (eds) Chapter 6 of Research tools in natural resource and environmental economics. World Scientific, Singapore, pp 187–213
- Graves PE, Waldman DW (1991) Multimarket amenity compensation and the behavior of the elderly. *Am Econ Rev* 81(5):1374–1381
- Graves PE, Murdoch JC, Thayer MA (1988) The robustness of hedonic price estimation: urban air quality. *Land Econ* 64(3):220–233
- Griliches Z (1961) Hedonic price indexes for automobiles: an econometric analysis of quality change. In: The price statistics of the federal government. NBER staff report no. 3, General series, no. 73. NBER, New York, pp 173–196
- Krumm R, Graves PE (1982) Morbidity and pollution. *J Environ Econ Manag* 9(4):311–327
- Palmquist RB (2005) Property value models. In: Maler K-G, Vincent JR (eds) Handbook of environmental economics: valuing environmental changes. North Holland, Amsterdam, pp 763–819, Vol. 2
- Ridker RG, Henning JA (1967) The determinants of property values with special reference to air pollution. *Rev Econ Stat* 49(2):246–257
- Roback J (1982) Wages, rents, and the quality of life. *J Political Econ* 90(6):1257–1278
- Rosen S (1974) Hedonic prices and implicit markets: product differentiation in pure competition. *J Political Econ* 82(1):34–55
- Taylor LO (2008) Theoretical foundations and empirical developments in hedonic modeling. In: Baranzini A, Ramirez J, Schraerer C, Thalmann P (eds) Hedonic methods in housing markets. Springer, New York, pp 15–38
- Waugh FW (1928) Quality factors influencing vegetable prices. *J Farm Econ* 10(2):185–196
- Zellner A, Siow A (1980) Posterior odds ratios for selected regression hypotheses. In: Bernardo JM, Degroot MH, Lindley DV, Smith AFM (eds) Bayesian statistics. Proceedings of the first international meeting held in Valencia. University of Valencia Press, Valencia, pp 585–603

Gara Villalba Méndez and Laura Talens Peiró

Contents

51.1	Introduction	1010
51.1.1	Material Balance by Total Mass	1011
51.1.2	Material Balance by Element	1011
51.2	Applications of Material Flow Analysis	1011
51.2.1	Studying Flows of Substances, Materials, and Products	1012
51.2.2	Studying Firms, Sectors, and Geographical Areas	1013
51.3	Case Studies	1015
51.3.1	Material Balance Applied to Chemical Industry	1015
51.3.2	Mass Balance Applied to Rare Earth Metals	1019
51.4	The Laws of Thermodynamics	1024
51.5	Conclusions	1026
	References	1026

Abstract

This chapter presents an overview of the mass balance principle and its applications. It is an important tool for quantifying wastes which are produced by economic processes. These wastes are equal in mass to the difference between total raw material inputs to the process and useful material outputs. Products are becoming more complex which results in an increase of input mass and wastes. It is safe to say that nowadays process wastes far exceed the mass of materials that are finally embodied in useful products.

G. Villalba Méndez (✉)
Universitat Autònoma de Barcelona, Bellaterra, Spain
e-mail: gara.villalba@uab.es

L. Talens Peiró
Social Innovation Centre, INSEAD, Fontainebleau, France
e-mail: laura.talenspeiro@insead.edu; laura.talens@gmail.com

The application of the mass balance principle can take many shapes and forms, and this chapter illustrates a few. Using mass balance and chemical engineering knowledge of processes, we found that on a yearly basis, the inorganic chemical industry has a yield of 91 % (9 % of the inputs end up in waste), and the organic chemical industry has a yield of 40 %. A second example is the rare earth metal industry, where potential recovery of these scarce metals is quantified to motivate reuse and recycling. Presently less than 1 % of rare earth metals are recovered from end-of-life products, but as the demand for these resources increases in the near future for products such as electric motors and wind power turbines, recovery will become necessary.

An introduction to thermodynamics and exergy is included, since all wastes are thermodynamically degraded as compared to raw materials. The exergy of the inputs, products, and wastes is an important factor to consider for process efficiency and environmental evaluation.

51.1 Introduction

We constantly perform material flow analysis in our daily activities without realizing it. For example, when we balance our checking account, we sum the money we are crediting to our account to the current balance and subtract our expenses. Unknowingly we are applying one of the most fundamental principles that govern our existence: the mass balance principle that states that *mass cannot be created nor destroyed*. This physical law has nontrivial consequences in economics. Economic processes require inputs, both energy and matter, and invariably generate waste. Economic products are becoming more and more complex requiring many times the materials and energy that are finally embodied in the product, resulting in many waste streams to land, air, and soil. A cellular phone 10 years ago required in the order of 20 different materials such as different plastics, copper, aluminum, and steel. Nowadays, a multifunctional mobile phone can have as many as one thousand different kinds of materials (Mueller et al. 2003). Eventually the products become waste themselves. This leads us to the idea of analyzing the *material life cycle*: a more comprehensive assessment of resource use and wastes also referred to as a “life cycle analysis” approach.

The following equation represents the mass balance principle:

$$\begin{array}{lclcl} \text{Mass} & & \text{Mass input} & & \text{Mass} \\ \text{accumulation} & = & \text{through} & - & \text{generation} \\ \text{within the} & & \text{system} & - & + \quad \text{within the} \\ \text{system} & & \text{boundaries} & & \text{system} \\ & & & & \text{Consumption} \\ & & & & - \quad \text{within the} \\ & & & & \text{system} \end{array}$$

The mass consumption and generation terms are associated to transformation due to chemical reactions. If there are no chemical reactions taking place, then these two terms are zero. If we assume steady state conditions, there is no accumulation of mass and the equation is further simplified.

There are two basic approaches that can be used to carry out material balance. The analysis can be performed based on (a) the total mass in each stream entering and leaving the system and/or (b) the composition of each stream entering/leaving the system. [Figure 51.1](#) represents the production of sulfuric acid (H_2SO_4) used to illustrate both approaches. In general terms, sulfuric acid production consists of a series of chemical reactions in which water, oxygen, hydrogen, and sulfur are needed and emissions such as SO_2 result as waste. Energy is needed in order to drive this process which consequently results in CO_2 emissions, useful work, and heat loss – but for simplicity we will not consider energy terms.

51.1.1 Material Balance by Total Mass

If we know the total amount (in kg) of sulfuric acid produced and the inputs required, we can easily calculate the amount of emissions that result from this process. This is illustrated by [Fig. 51.1a](#) for the production of 1 kg of H_2SO_4 and Eq. (51.1) applied to this case:

$$\begin{aligned} \text{Total input} &= \text{product} + \text{waste} \\ (190 \text{ g } H_2O + 500 \text{ g } O_2 + 330 \text{ g } S) - 1000 \text{ g } H_2SO_4 &= \text{emissions} = 20 \text{ g emissions} \end{aligned} \quad (51.1)$$

51.1.2 Material Balance by Element

If we want to know the composition of emissions, we could perform a type (b) analysis where a mass balance is performed element by element. Since the composition of the product is known, we can calculate the composition of the waste using the molecular weight (MW) of each element. This is illustrated in [Fig. 51.1b](#): there is an input of 670 g of O which comes from the water and O_2 . We also know that the product output is 1 kg of H_2SO_4 which is equivalent to 10.2 moles of H_2SO_4 , 40.2 moles of O, or 650 g of O. Applying Eq. (51.2) by element gives the following:

$$\begin{aligned} \text{Total O input} &= \text{O in product} + \text{O in waste} \\ 670 \text{ g O} - 650 \text{ g O in product} &= \text{O in waste} = 10 \text{ g O} \end{aligned} \quad (51.2)$$

So now we know that of the 20 g of emissions calculated using approach (a), 10 g is of oxygen. This is just based on some simple calculations, but if we were to use simulation software that also takes thermodynamics into consideration, we could in theory know in what compound and in what state that oxygen ends up as in the waste stream.

To summarize, the MFA procedure follows these steps:

1. Define the process under study and the system boundaries, both spatially and temporally.
2. Label all flows, inputs, outputs, and accumulation.

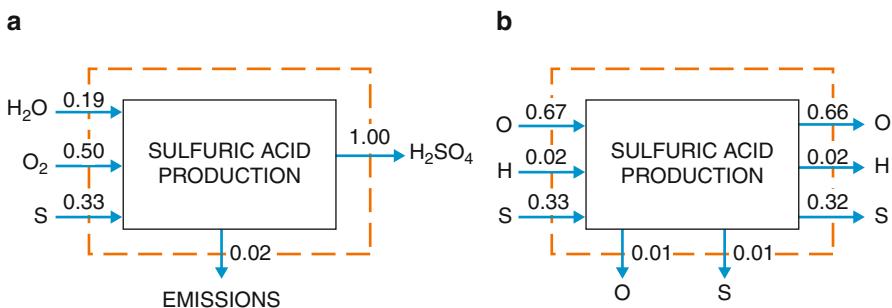


Fig. 51.1 Sulfuric acid production: (a) mass balance by chemical compound and (b) mass balance by chemical element

3. Identify all known values of composition and stream flows.
4. List all independent mass balances that are possible. Sometimes assumptions must be made when not sufficient data is available.
5. Solve the equations for unknown variables.

51.2 Applications of Material Flow Analysis

The field of industrial ecology is well known for evaluating industrial systems based on material flow analysis. The industrial system under study can be at any level: the most encompassing one being global and the simplest being a single manufacturing process such as the sulfuric acid production illustrated earlier. The basic approach is that the system to be analyzed is viewed as a transformation process that requires certain inputs such as material, energy, and “free goods” from the environment. These are converted to products, by-products, and wastes that can be airborne, liquid, or solid. In other words, the system “digests” raw materials into products and the whole process is referred to as “industrial metabolism.” Thanks to the mass balance principle, wastes can be calculated if we know the sum of the inputs and useful outputs.

There are different types of MFA depending on what system needs to be evaluated and what the objectives are. Figure 51.2 summarizes the types of material flow-related analysis that can be done. Basically, MFA is divided into two types: (a) for studying a specific environmental problem related to certain impacts per unit of flow of substances, materials, and products within certain firms, sectors, and regions and (b) for analyzing problems of environmental concern related to the throughput of firms, sectors, and regions associated with substances, materials, and products.

51.2.1 Studying Flows of Substances, Materials, and Products

The main purpose of studying flows of substances, materials, and products through a system is to clearly define the metabolism of the material or system under study.

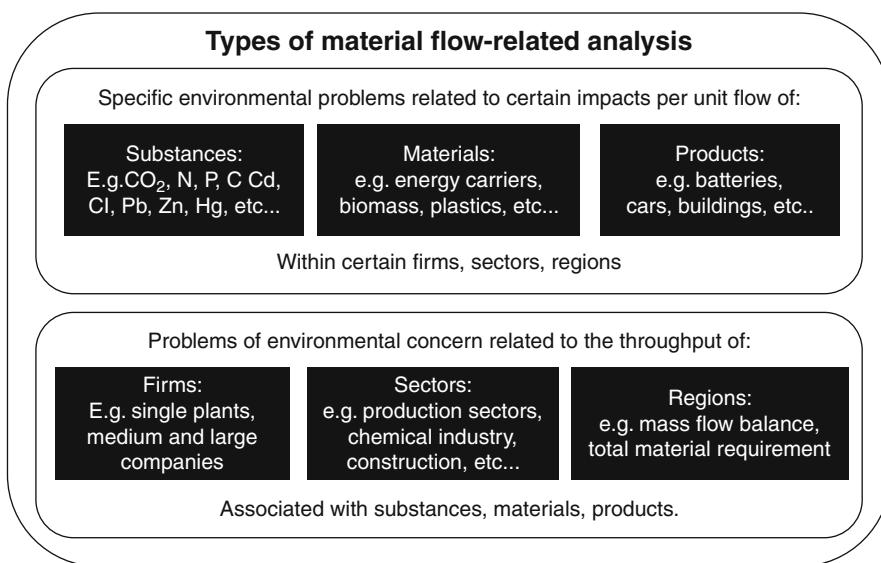


Fig. 51.2 Types of material flow-related analysis (Source: Adapted from Bringezu and Moriguchi (2002))

For example, the study of major chlorine-based chemicals defines the chlorine metabolism throughout the economy and also identifies other production processes where chlorine (Cl₂) is used to produce intermediate products such as caustic soda (NaCl) or final products such as paper (Ayres and Ayres 1998). Figure 51.3 shows the material flows in the production of major chlorine-based chemicals in the USA in the year 1993. This study quantifies all chlorine wastes and emissions during production and also identifies the most intensive chlorine-consuming sectors, in this case the organic synthetics (11.15 million t) and paper and pulp mills (4.14 million t). It was useful to quantify chlorine losses because of its toxic potential and various pollution problems, the ozone depleting effect of chlorofluorocarbons (CFCs) and the risks incurred through the incineration of materials such as polyvinylchloride (PVC). This type of MFA, also referred to as substance flow analysis (SFA), determines the main entrance routes of chlorine to industry which is useful for qualitatively assessing risks to substance-specific endpoints (Van der Voet 2002).

51.2.2 Studying Firms, Sectors, and Geographical Areas

MFAs can also be applied to firms, sectors, or geographical areas, to evaluate their environmental performance. For example, by performing an MFA to a firm producing chlorine (Cl₂), we can calculate material requirement for its production (resource depletion) and the wastes and emissions per tonne of chlorine produced. Figure 51.4 illustrates these figures for the production of 1 t of chlorine (Cl₂) by

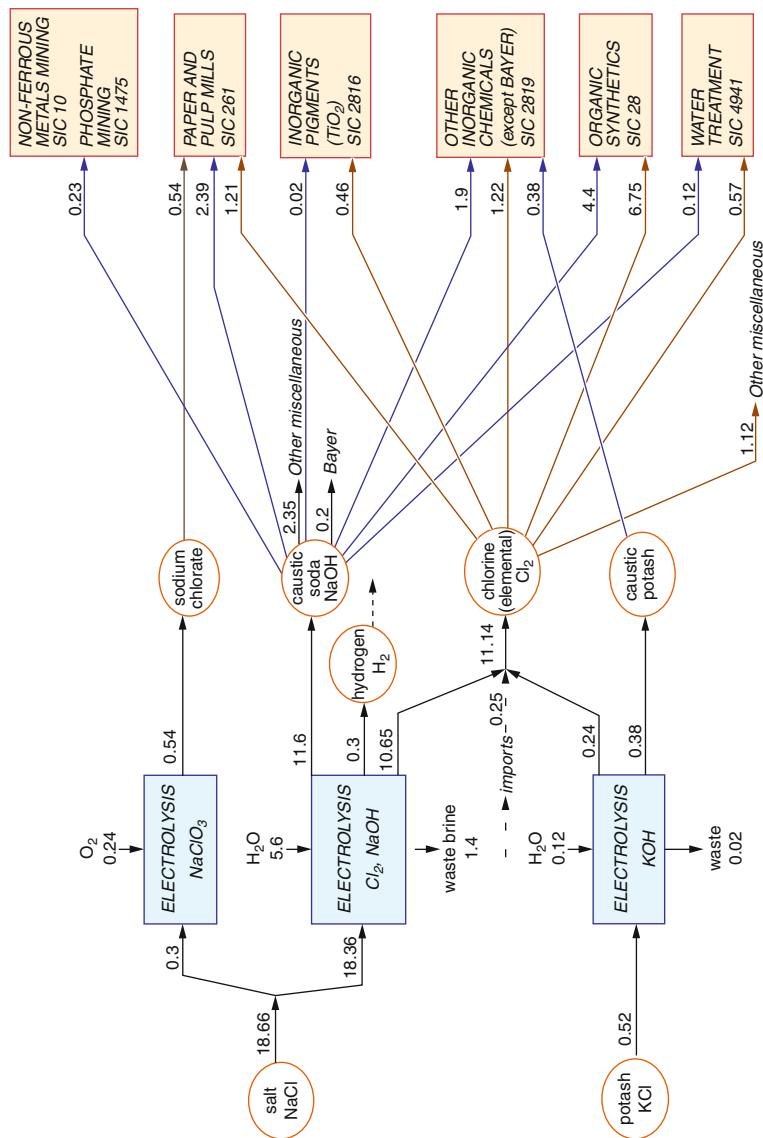


Fig. 51.3 Material flows in the production of major chlorine-based chemicals in the USA, 1993 (tonnes) (Source: Ayres and Ayres (1998))

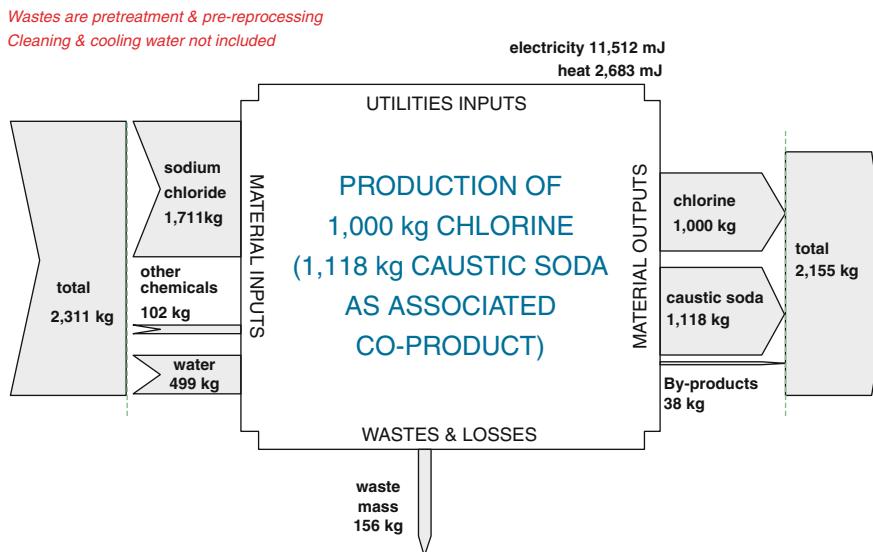


Fig. 51.4 Chlorine by electrolysis of sodium hydroxide (NaOH) in mercury cells (Source: Adapted from Ayres and Ayres (1998))

electrolysis of sodium chloride (NaCl). The inputs for this process are as follows: 1,711 kg of sodium chloride (NaCl), 499 kg of water (H_2O), and 102 kg of other chemicals which are not specified. As a result, 156 kg of waste is produced.

When applying MFA to a product or firm, a life cycle approach is normally taken and is denoted life cycle assessment (LCA). An LCA accounts the inflows and outflows of the system “from cradle to grave”; that includes all material inputs and outputs from extraction, manufacturing, consumption or use, recycling, and final disposal. The initial interest in developing LCA was to minimize the energy consumption and solve the waste management problems. The first LCA project, originally called REPA (Resource and Environmental Profile Analysis), was carried out by the Midwest Research Institute for the Coca-Cola Company in 1969. The goal was to compare several container options by quantifying emissions, material, and energy consumption of each. Presently, LCA is standardized by the International Standard Organization on the ISO14,040 series and Life Cycle Initiative program led by the United Nations Environment Programme (UNEP) and the Society for Environmental Toxicology and Chemistry (SETAC) created to develop and disseminate practical tools for evaluating the opportunities, risks, and trade-offs, associated with products and services over their whole life cycle (Mila i Canals 2003).

Section 51.3 illustrates different MFA approaches. First in Sect. 51.3.1, an analysis of the inorganic and organic chemical industry is given to quantify wastes and process conversion. Section 51.3.2 shows a material flow analysis of rare earth metals by current market demand. This is useful in order to quantify potential recovery of these critical metals in the waste streams.

51.3 Case Studies

51.3.1 Material Balance Applied to Chemical Industry

Quantitative data about industrial chemicals can be estimated with reasonable accuracy from industry production statistics. In the USA, production statistics were published annually by the US International Trade Commission (USITC) until the mid-1990s. The USITC reports included production data for virtually all industrial chemicals, including intermediates. Hence, in this example, we use production statistics from USITC for years 1991–1993. Unfortunately, these reports are no longer published.

To compare inputs and outputs for the whole sector and avoid double counting, the list can be divided into two groups: (i) basic chemicals, which are made directly from raw materials, and (ii) all others, including intermediates. For such classification, some knowledge of the industry is required. For instance, sulfuric acid is mainly made by burning sulfur but is now also produced as a by-product of copper smelting. Hydrochloric acid is no longer made from salt but as by-product of many downstream chlorination processes and thus is not considered a “basic” chemical.

Based on process information of the basic inorganic and organic chemicals, raw material inputs are quantified in mass terms, whence a material balance by elements (C, H, O, N, Cl, S, Na, Ca, etc.) can be performed. The difference between mass inputs and useful outputs is wastes and emissions. For the industry as a whole, wastes are characterized by elemental composition, but they can be estimated approximately as a mix of compounds (CO_2 , CO, H_2O , NaCl, CaSO_4 , etc.) based on knowledge of process reactions.

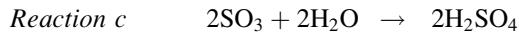
51.3.1.1 Inorganic Chemicals

Based on process information, the basic inorganic chemicals included are sulfuric acid, ammonia, chlorine, and caustic soda. The total production of these four chemicals represents 75 % of the total mass of inorganic chemical production (Ayres and Ayres 1999). Once the outputs are identified, we need to estimate the inputs. Mass inputs are identified based on the theoretical reaction for the production of each inorganic product.

Sulfuric Acid

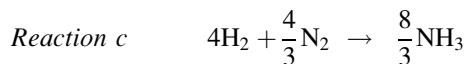
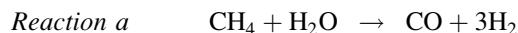
Sulfuric acid can be produced from a large number of raw materials: crude oil and natural gas, copper and lead-zinc ores, organic spent acids, sulfur-containing gases, and sulfur salts. However, in practice, it is mainly produced from sulfur, oxygen, and water by single-/double-contact absorption – when sulfur has been purified and dried – and wet/combined dry-wet catalysis – when sulfur originates from the burning or the catalytic conversion of hydrogen sulfide (H_2S) gases – as illustrated in the three reactions below:





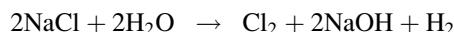
Ammonia

Ammonia is a starting material used in a wide variety of industrial chemicals and nitrogen fertilizers. The latter are responsible for 90 % of all ammonia production to obtain urea, ammonium nitrate, ammonium sulfate, and ammonium phosphates (Suresh and Fujita 2007). Most of the world's ammonia is produced from natural gas by steam reforming, except in China where ammonia is produced from synthesizing gas from coal. Steam reforming involves two main reactions: the separation of hydrogen from methane (reaction a) and its recombination with atmospheric nitrogen (reaction c):



Chlorine and Sodium Hydroxide

Chlorine and sodium hydroxide are produced as coproducts by electrolytic decomposition of sodium chloride solutions obtained from brines. The electrolysis of chlorine consists on using direct electric current to drive chemical reactions, in this case to dissociate sodium chloride in sodium cations and chlorine anions (Bommaraju et al. 2000). During the electrolysis process, chlorine anions are oxidized at the anode to produce chlorine and sodium cations with hydroxyl anions from water form sodium hydroxide at the cathode. Besides chlorine and sodium hydroxide, hydrogen is also generated (as illustrated in the reaction below):



Results

The overall mass inputs for the production of the main inorganic chemicals are hydrogen, methane, nitrogen, oxygen, sodium chloride, and sulfur. Figure 51.5 shows the elemental and component mass balance for the production of basic inorganic chemicals in the USA in 1991. Performing a mass balance by elements and components helps ensure the consistency of inputs and outputs. Mass inputs are estimated based on the production statistics of end-products and the reactions illustrated above. The mass balance shows that about 9 % of the total mass inputs are

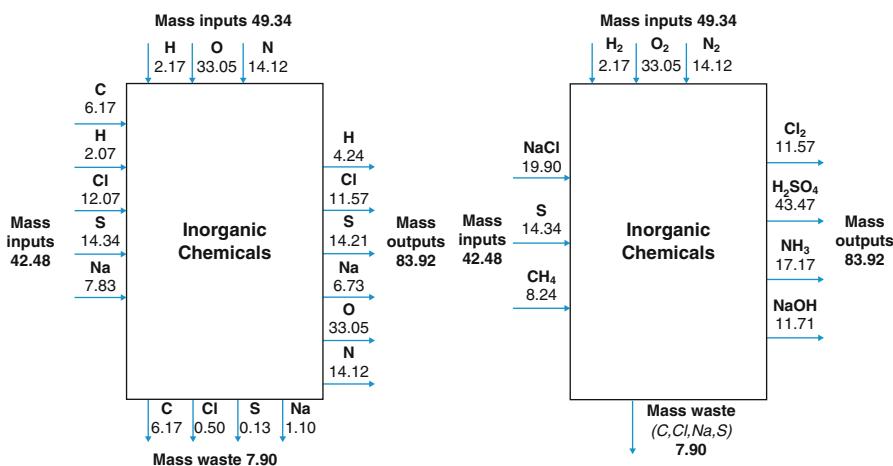


Fig. 51.5 Mass balance by elements and compounds of the production of inorganic chemicals in 1991 (MMT) (Source: Ayres et al. (2011))

wasted as compounds made of carbon, chlorine, sodium, and sulfur. The process conversion or yield can be estimated by dividing the mass output of the products by that of the inputs. For inorganic chemicals, such conversion equals 91 %.

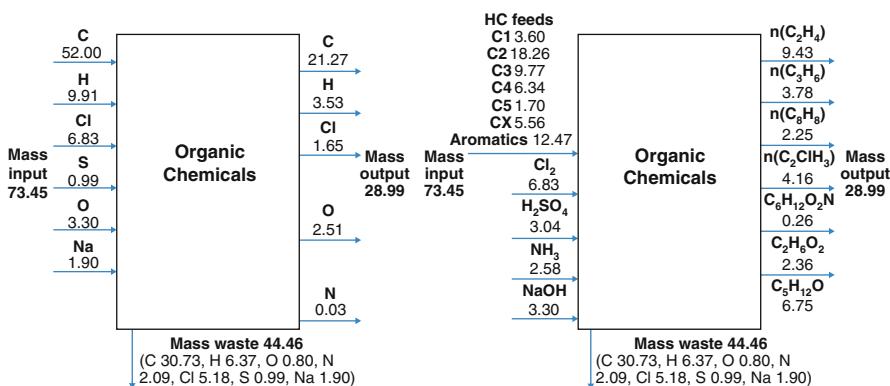
51.3.1.2 Organic Chemicals

Based on USITC statistics for 1991–1993, the major end products from organic chemicals are plastics (as polyethylene, polypropylene, polystyrene, and polyvinyl chloride), nylon 6, ethylene glycol (antifreeze), and methyl tert-butyl-ether (A fuel additive that has been largely phased out since 1991). The production of these chemicals, in mass terms, represented about 80 % of total US production of organics in 1991 (Ayres and Ayres 1998). Basic organic inputs are hydrocarbons and also inorganic raw materials as chlorine, sulfuric acid, ammonia, and caustic soda. Most organic chemicals are produced from feedstocks from natural gas or petroleum refineries, with a very small share from coal. There are three categories of feedstock: paraffin, olefins, and cyclic/aromatics. Paraffins are saturated straight- or branched-chain hydrocarbons. Examples include methane, ethane, propane, isobutene, and n-butane. Olefins are unsaturated aliphatic compounds with one or more double bonds. Examples include ethylene, propylene, butylenes, and butadiene. Cyclic aromatics are benzene, toluene, xylene, cyclopentene, cyclohexane, and naphthalene. Table 51.1 shows the organic chemicals and their primary production.

The mass input of the inorganic raw materials chlorine, sulfuric acid, ammonia, and caustic soda is estimated based on their use patterns (Ayres and Ayres 1998). Figure 51.6 illustrates material balance by elements and components for the production of the listed basic organic chemicals. The mass balance shows that about 60 % of the total mass inputs are wasted as compounds of carbon, hydrogen, oxygen, nitrogen, chlorine, sodium, and sulfur. The process conversion or yield for organic chemicals is about 40 %.

Table 51.1 US primary feedstock production in 1991 (MMT) (Source: Ayres and Ayres 1998)

Organic chemicals	Formula	Mass (MMT)
<i>Aliphatics and olefins</i>		
C ₁ Methane	CH ₄	3.60
C ₂ Acetylene	C ₂ H ₂	0.14
Ethylene	C ₂ H ₄	18.12
C ₃ Propylene	C ₃ H ₆	9.77
C ₄ Butylene	C ₄ H ₈	1.05
Butadiene	C ₄ H ₆	1.39
Butene	C ₄ H ₈	0.43
Isobutane	C ₄ H ₁₀	0.50
Isobutylene	C ₄ H ₈	0.44
Other C ₄	C ₄ H ₈	2.54
C ₅ Isoprene	C ₅ H ₈	0.21
Pentene, mixed	C ₅ H ₁₀	0.19
Other C ₅		1.29
C _x All other aliphatics (including methane)		5.56
<i>Aromatics and naphthalenes</i>		
Benzene, all grades	C ₆ H ₆	5.21
Toluene, all grades	C ₇ H ₈	2.86
Xylenes, all grades	C ₈ H ₁₀	2.87
All other aromatics and naphthalenes		1.54
<i>Total inputs</i>		57.70

**Fig. 51.6** Mass balance by elements and compounds of the production of organic chemicals in 1991 (MMT) (Source: Ayres et al. (2011))

51.3.2 Mass Balance Applied to Rare Earth Metals

A material flow analysis (MFA) helps determine the main entrance routes of rare earth (RE) metals to the economic system. Such quantification is useful to quantify

future potential recovery of these critical metals in end products once they reach their end of life. The amount of RE metals in intermediate and end products can be estimated based on the production and market share of each metals (Kingsnorth 2009; Chegwidden and Kingsnorth 2010; Morgan 2010; Schüler et al. 2011). The main functions of RE are as dopants for semiconductors, catalysis, electricity storage, alloying elements, additive in glass and ceramic, and as abrasives. All these functions are used in intermediates and end products. For example, lanthanum is used as dopant for semiconductors in phosphors (intermediate product), and phosphors are used in liquid crystal display, plasma flat panels, and lighting, all of them end products.

51.3.2.1 Dopants for Semiconductor in Phosphors

The principal applications for RE phosphors are in display screens (cathode ray, liquid crystal, and plasma) and in low-energy fluorescent lighting tubes. Each of the different display technologies requires different types and compositions of phosphors, as do fluorescent tubes in which the phosphors reduce energy consumption and provide specific colors. Phosphors consist of a host material with an added activator or dopant. For red, the RE oxides used include yttrium, europium, and gadolinium. For green, the hosts used are lanthanum, cerium, and yttrium, while terbium and gadolinium are used as activators. For blue, europium oxide is mainly used. The combination of red, green, and blue gives white color. In 2010, 8,250 t of RE was used to produce phosphors: 6,135 t for red, 2,065 t for green, and 50 t for blue color. Phosphors are largely used for lighting (84 %), followed by LCDs (12 %) and plasma displays (4 %).

51.3.2.2 Catalysts

In 2010, 22,920 t of RE was used as catalyst. About 70 % was used in fluid catalytic cracking (FCC) and 30 % in autocatalyst converters. Cerium and neodymium are also used in non-cracking catalyst processes such as ammonia synthesis, hydrogenation, dehydrogenation, polymerization, isomerization, and oxidation and in automobile emissions control; however, the amounts used for this purposes are not published.

Fluid Catalytic Cracking (FCC)

FCC is mainly used in petroleum refining to break down long complex organic molecules as kerosene and heavy hydrocarbons into simpler and lighter molecules as gasoline and liquefied petroleum gas. The most widely used catalysts are synthetic zeolites (zeolite Y and ZMS-5 zeolites) which contain lanthanum and cerium that improves the stability at high temperature and increases catalyst activity and gasoline selectivity (Yang et al. 2003). Commercial catalysts are composed by 85 % amorphous silica-alumina cracking catalyst and 15 % of zeolites with a varying content of 0.2–3 % of RE (Estevao et al. 2005; Xiaoning et al. 2007; Schiller 2011). In 2010, the annual production of feedstock from FCC was 1,668 million l. FCC containing primarily Y zeolites account for more than 95 % of total

consumption (Davis and Inoguchi 2009). Assuming that production of RE for FCC requires 15,940 t of RE, for each liter of feedstock, an average of 9 g of RE is required, that is, 0.2 % of RE content.

Autocatalyst Converters

RE metals are also key for autocatalyst converters to reduce the emission of carbon monoxide (CO), hydrocarbons (HC), and nitrogen oxides (NO_x). They are added to the wash-coating to improve the thermostability of alumina and ensure the activation of catalyst under high temperature. In 2010, the manufacturing of 78 million units of cars required a total of 6,980 t of RE, which gives an estimate of 90 g of RE per vehicle, one-third of the amount reported in 2003 (Xiaodong and Duan 2004). The composition of RE in converters is 90 % cerium, 5 % lanthanum, 3 % neodymium, and 2 % praseodymium. Thus, in 2010, 6,280 t of cerium, 350 t of lanthanum, 210 t of neodymium, and 140 t of praseodymium were used in internal combustion vehicles.

51.3.2.3 Electrical Storage in NiMH Batteries

A nickel-metal hydride (NiMH) battery is a type of rechargeable battery composed by cathode, anode, electrolyte, and separator, all assembled in a steel case. Xu estimated the following content of metals: 50 % nickel, 33 % RE, 10 % cobalt, 2 % aluminum, and 6 % manganese (Xu and Peng 2009). RE metals are mainly contained in the anode of NiMH batteries which are described as AB_5 where A stands for lanthanide metal and B for nickel. In practice, lanthanum is substituted by lanthanum-rich mischmetal containing 50 % lanthanum, 33 % cerium, 3 % neodymium, 10 % praseodymium, and 3 % samarium (Morgan 2010). In 2010, 12,670 t of RE was used in NiMH battery alloys. For HEV which represents 65 % of the end use of NiMH batteries, the total amount of RE equals 8,060 t. According to Pillot, the remaining 4,610 t was used for retail (toys and household tools), cordless phones, and other electric and electronic devices (Pillot 2011).

51.3.2.4 Alloying Elements

Almost all the published estimation of the amount of RE metals for metallurgy agrees on giving an estimate of 32,025 t (Kingsnorth 2009; Chegwidden and Kingsnorth 2010). From this production, 75 % are used in magnets and 25 % in alloys with iron and aluminum. It is assumed that the usage of RE in magnets corresponds to the composition in neodymium-iron-boron (NIB) magnets whose average amount is 30 % RE, 69 % iron, and 1 % boron (Morgan 2010). For 2010, we estimated a total of 24,060 t used in NIB magnets. Magnets are used in wind turbines, hybrid vehicles, magnetic resonance imaging (MRI), and electric and electronic devices.

Wind Turbines

In 2010, the new wind turbine installation was 36 GW, and only about 14 % of the total new installation used NIB magnets (Schüler et al. 2011). Each MW of wind

turbine installed requires 860 kg of NIB magnets. Based on the composition given by Morgan, 910 t of neodymium, 310 t of praseodymium, 70 t of dysprosium, and 10 t of terbium were used for NIB magnets in wind turbines.

Electric Vehicles

The number of new hybrid cars registered reached 533,000 units in 2010. Assuming that each electric vehicle requires 1 electric motor per wheel and that the average amount of neodymium per electric vehicle is 6.3 kg, the total amount of RE is 4,800 t (Talens Peiró et al. 2013). The amount of each RE metal is 3,358 t of neodymium, 1,148 t of praseodymium, 264 t of dysprosium, and 34 t of terbium.

Magnetic Resonance Imaging (MRI)

In 2010, 2,500 new MRI units each of them using an average of 860 kg of NIB magnets were produced (Cosmus and Parizh 2011). The amount of RE required by them was 450 t of neodymium, 5 t of terbium, 155 t of praseodymium, and 35 t of dysprosium.

Gadolinium is a minor RE metal used in the magnet sector as an MRI contrast agent and as magnet component in research for magnetic cooling. As an MRI agent, it improves the visibility of internal body structures by altering the relaxation times of tissues and body cavities where it is present. Gadolinium is used in doses of about 0.01–0.03 g per kg of body mass (Niendorf et al. 1991). For the 80 million MRI exams performed in 2010, 90 t of gadolinium was used (Cosmus and Parizh 2011). Gadolinium is also used in magnetic cooling research as powder for creating magnetic refrigeration. In 2010, about 390 t was used for this purpose.

Minor Alloys

RE mischmetal is also used as minor alloys for controlling inclusions and improving the performance for steel and iron. For instance, cerium combined with sulfide forms particles more rounded that are less likely to generate cracking. RE mischmetal is used in zinc galvanizing applications as for zinc-aluminum alloy named Galfan (Zn-5Al-MM) which is often used as the coatings for steel, to enhance the product life for certain applications. In 2010, 7,965 t of RE was used as mischmetal in iron and aluminum alloys.

51.3.2.5 Additives

Additives are substances added to preserve the quality and appearance of coatings and for coloring. They are widely used in the glass and ceramic industry for quality and as colorants. In 2010, the total amount of REE used as additives was 17,425 t: 37 % in ceramic and 63 % in glass.

Glass Industry

Cerium and lanthanum oxides are used in glass to overcome the decolorizing to yellow green caused by iron oxide, always present as an impurity glass. Cerium is also a good UV and IR absorbent and thus used in protective glasses and in quantities of 2–4 % for glass blowing and welding goggles (Gupta and

Krishnamurthy 1992). Lanthanum is used in silica glasses to give a high index of refraction and low dispersion in lenses for autofocus single-lens reflex (SLR) cameras and video cameras. Other REE used in lower amounts are neodymium, yttrium, and praseodymium. Neodymium and praseodymium are used for coloring glasses. Neodymium colors glass bright red, praseodymium colors glass green, and their combination colors blue. Yttrium is used in the form of yttrium-aluminum garnets ($\text{Y}_3\text{Al}_5\text{O}_{12}$) to form synthetic crystals that are widely used as an active laser medium in solid-state lasers. YAG lasers use neodymium for its optimal absorption and emitting wavelength to be used in various medical applications, drilling, welding, and material processing. Other metals used as additives are erbium, ytterbium, and holmium which are used in luminescent solar concentrators and light sources for fiber optics and in laser materials.

Ceramic Industry

In 2010, 6,865 t of RE was used by the ceramic industry. The RE used were yttrium (3,495 t), lanthanum (1,190 t), cerium (980 t), neodymium (800 t), and praseodymium (400 t). Yttrium is used combined with silica for turbine blade applications. Cerium is used as phase stabilizer in zirconia-based products and in various ceramics including dental compositions. Yttrium and cerium are used in partially stabilized zirconia (PSZ) and tetragonal zirconia polycrystals (TZP), both high-performance ceramics with excellent toughness and strength properties at low and intermediate temperatures. End products of these ceramics containing yttrium are components for adiabatic diesel engines, cutting tools, wire drawing dies, and furnace elements for use up to 2,000 °C in oxidizing atmosphere. Lanthanum is used in lead-zirconate-titanate (PLZT), a transparent ferroelectric ceramic material. Neodymium is used in ceramic glazes to produce blue to lavender colors. Praseodymium incorporated in zirconium silicate lattice is used for the production of high temperature resistance lemon yellow pigments for the ceramic industry.

51.3.2.6 Abrasive

RE oxides are excellent abrasives for glass polishing in the manufacture of LCD, optical glass, mirrors, photomasks, plate glass, lenses, and cut glass. RE oxide powders provide a high mechanical abrasion react with the surface of glasses plus a high-quality finishing (Xu and Peng 2009). There are various grades of RE oxide polishing powder. They can be fully composed by cerium oxide, or with a content of 45–75 % cerium with the remaining of other RE oxides. The average composition of polishing powders is 32 % lanthanum, 65 % cerium, and 4 % of praseodymium (Morgan 2010). In 2010, 13,750 t of REE was used as polishing powder in the glass industry, about 40 % of which was consumed in LCD industry.

51.3.2.7 Results of MFA of Rare Earths

With each of rare earth metals identified and quantified in intermediate products, we can also estimate their content in end products. This quantification helps do some estimation about their recovery. Not all of these metals can be recovered in practice; however, if we know what the intermediate and end products are used for,

a theoretical potential recovery can be calculated for non-dissipative uses. For example, let us trace the 24,060 t of RE in magnets in 2010, which lie under the function “alloying elements” in Fig. 51.7. Magnets are used in wind turbines, electric vehicle batteries, magnetic resonance imaging (MRI), electronic products, and magnetic cooling applications. If we know the amount of each metal present in each of these end products, we could calculate what could, in theory, be recovered at the end of life. Following our example, based on several references and estimations, we calculated that 1,300 t of RE was embodied in magnets of wind turbines, of which 910 t was of neodymium. Using the same approach, we found that 3,358 t of neodymium was used in magnets in electric vehicles, 450 t in MRI, and 11,980 t in electric and electronic devices, totaling 16,700 t of neodymium that could in theory be recovered at some point in the future. This is a substantial amount if we compare with neodymium production for that year which was 21,615 t. According to Graedel, the actual end-of-life recycling rate for neodymium is less than 1 % (Graedel 2011).

If we look at phosphor applications, in 2010 a total of 8,250 t of REM was consumed. These ended up in lighting applications, liquid crystal display screens (LCD), and plasma panels. If we trace europium which is the only metal known that can emit blue light, 24 t was used in lighting, 20 in LCD, and 6 in plasma panels, adding to a total of 50 t of europium used for blue phosphors. Graedel estimates that less than 1 % of europium is recycled at the end of life of these products.

As the demand of rare earth metals increases in the future, it will become necessary to increase the recovery of these metals at the end of life of the products that contain them and improve present production from base metals. MFA helps identify potential sources.

51.4 The Laws of Thermodynamics

Material flow analysis is useful in quantifying resource consumption, wastes, and process losses. However, it becomes even more useful when we combine it with the first and second laws of thermodynamics. Energy, just like matter, cannot be created nor destroyed; energy is conserved in every action or transaction. That is the first law of thermodynamics, perhaps the most fundamental of all physical laws. But energy can be degraded and be transformed to “less useful” types of energy such as low-grade heat. This fact is a consequence of the second law of thermodynamics, sometimes known as the entropy law that states that global entropy increases in every irreversible process.

Exergy is measure of the potential work that can be performed by a system (Szargut et al. 1988). In other words, as a system degrades, its entropy increases and its exergy, or potential to do work, decreases. Exergy is not conserved: the exergy component is “used up” as it does work.

Exergy is also a thermodynamic quantity that reflects the “distance” from thermodynamic equilibrium of a “target” material, or subsystem. It is therefore

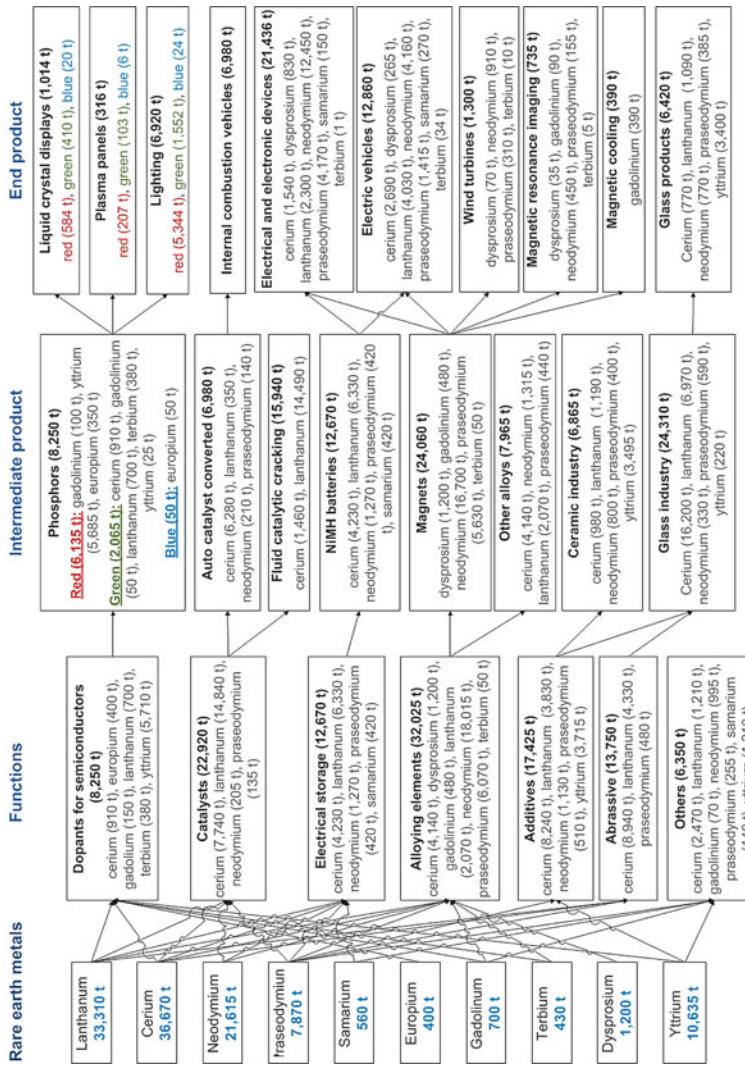


Fig. 51.7 Rare earth metals contained in intermediate and end products (Source: Talens Peiró et al. 2012)

only definable in terms of a reference state, such as the atmosphere, the ocean, or the average earth's crust, depending on which of them the subsystem will eventually rejoin and become indistinguishable from. Since the quantity of exergy contained in a subsystem is a measure of potential work, it is measurable in the same units as energy and work (joules, kWh, Btu, etc.). In the case of fuels, the exergy content is almost exactly the same as the enthalpy or heat content. For foods, the exergy content is essentially the same as the calorie content. For metals, the exergy content is the amount of heat that would be generated if the metal were to be completely oxidized. Exergy has been calculated and tabulated in reference books such as Szargut et al. (1988).

Calculating all flows in exergetic terms allows us to include both material and energy in our balance model. This is especially useful in order to calculate process efficiency and resource productivity. For more literature on exergy and material flow analysis, see Chapter "Material Balance Models in Research Tools in Natural Resources" in Ayres and Villalba Méndez (2011).

51.5 Conclusions

Material balance is based on the mass conservation principle which states that the sum of the weight of all inputs must be exactly equal to the sum of all outputs. Such simple postulate provides significant information when used for evaluating systems. First, when inputs and outputs are known, amount of wastes can be calculated, and the composition can be identified when chemical reactions. Second, it helps estimate the conversion or yield of processes which serves as a measure of the efficiency of the process in mass terms. There are many possible applications depending on one objective. For example, material balances across a geographical level aid to monitor the physical flows of materials across regional boundaries. At global level, MFA can be used to identify the end uses of materials which serve to estimate potential recovery.

Decision makers, engineers, and researchers need appropriate information tools to evaluate resource intensity, processing technologies, and resource efficiency. Using material flow analysis to evaluate economic processes helps identify opportunitiest strategies. For reducing waste and increasing recovery, and recycling. The combination of material balance and exergy analysis is the next logical step to accounting wastes and emissions since exergy gives a quantitative and qualitative measure of material potential usefulness.

References

- Ayres RU, Ayres LW (1998) Accounting for resources 1: economy-wide applications of mass-balance principles to materials and waste. Edward Elgar, Cheltenham/Lyme
Ayres RU, Ayres LW (1999) Accounting for resources 2: the life cycle of materials. Edward Elgar, Cheltenham/Lyme

- Ayres RU, Villalba Méndez G (2011) Materials balance models book section materials balance models. In: Batabyal AA, Nijkamp P (eds) Research tools in natural resource and environmental economics. World Scientific, Singapore, pp 403–422
- Ayres RU, Talens Peiró L, Villalba Méndez G (2011) Exergy efficiency in industry: where do we stand? *Environ Sci Technol* 45(24):10634–10641
- Bommaraju TV, Lüke B, O'Brien TF, Blackburn MC (2000) Chlorine. Book section chlorine (5 th ed): Kirk-Othmer Encyclopedia of chemical technology. Wiley & Sons, New York
- Bringezu S, Moriguchi Y (2002) Material flow analysis. Book section material flow analysis. In: Ayres RU, Ayres LW (eds) A handbook of industrial ecology. Edward Elgar, Cheltenham/Lyme, pp 79–90
- Chegwidden J, Kingsnorth D (2010) Rare earths – a golden future or overhyped? 20th industrial minerals international congress and exhibition, Miami
- Cosmus T, Parizh M (2011) Advances in whole-body MRI magnets. *IEEE Trans Appl Supercond* 21(3):2104–2109
- Davis S, Inoguchi Y (2009) Chemical economics handbook. Stanford Research Institute, Stanford, CA
- Estevao LR, Le Bras M, Delobel R, Nascimento RSV (2005) Spent refinery catalyst as a synergistic agent in intumescence formulations: influence of the catalyst's particle size and constituents. *Polym Degrad Stab* 88(3):444–455
- Graedel TE (2011) On the future availability of the energy metals. *Ann Rev Mater Res* 41(1):323–335
- Gupta CK, Krishnamurthy N (1992) Extractive metallurgy of rare earth. *Int Mater Rev* 37(5):197–248
- Kingsnorth D (2009) The rare earths market: can supply meet demand in 2014? Prospectors and developers association of Canada, Toronto
- Mila i Canals L (2003) Contributions to LCA Methodology for agricultural systems. Institut de Ciència i Tecnologia Ambientals ICTA. Bellaterra, Barcelona, Universitat Autònoma de Barcelona, p. 250
- Morgan JP (2010) Rare earths. We touch them everyday. Australia Corporate Access Days, New York
- Mueller J, Griese H, Hageluken M, Middendorf A, Reichl H (2003) X-free mobile electronics-strategy for sustainable development. IEEE international symposium on electronics and the environment. Vienna, pp 13–18
- Niendorf HP, Haustein J, Cornelius I, Alhassan A, Clauß W (1991) Safety of gadolinium-DTPA: extended clinical experience. *Magn Reson Med* 22(2):222–228
- Pillot C (2011) HEV, P-HEV and EC market 2010–2020. Impact on the battery business. 4th International congress on automotive battery technology, Wiesbaden
- Schiller R (2011) Optimizing FCC operations in a High Rare Earth Cost Market: Part I, available at online journal: Refinery Operations, August 3, 2011, vol2, issue 15 pp 1-2, accessed on January 20th 2013 at http://refineryoperations.com/downloads/refinery-operations_2-15_2011-08-03.pdf
- Schüler D, Buchert M, Liu R, Dittrich S, Merz C, Merz C (2011) Study on rare earths and their recycling. Öko-Institut e.V, Darmstadt, p 162
- Suresh B, Fujita K (2007) Ammonia. Stanford Research Institute
- Szargut J, Morris DR, Steward FR (1988) Exergy analysis of thermal, chemical, and metallurgical processes. Hemisphere Publishing Corporation, New York
- Talens Peiró L, Villalba Méndez G, Ayres RU (2013) Material flow analysis of scarce metals: sources, functions, end-uses and aspects for future supply. Environmental Science and Technology, Accepted for publication (2013)
- Van der Voet E (2002) Substance flow analysis (SFA) methodology. Book section substance flow analysis (SFA) methodology. In: Ayres RU, Ayres LW (eds) A handbook of industrial ecology. Edward Elgar, Cheltenham/Lyme, pp 91–101

- Xiaodong W, Duan W (2004) Development of auto exhaust catalysts and associated application of rare earth in China. *J Rare Earths* 22(6):837–843
- Xiaoning W, Zhen Z, Chunming X, Aijun D, Li Z, Guiyuan J (2007) Effects of light rare earth on acidity and catalytic performance of HZSM-5 zeolite for catalytic cracking of butane to light olefins. *J Rare Earths* 25(3):321–328
- Xu T, Peng H (2009) Formation cause, composition analysis and comprehensive utilization of rare earth solid wastes. *J Rare Earths* 27(6):1096–1102
- Yang H, Wang H, Yu H, Xi J, Cui R, Chen G (2003) Status of photovoltaic industry in China. *Energy Policy* 31(8):703–707

Amy W. Ando and Kathy Baylis

Contents

52.1	Introduction	1030
52.2	Spatial Heterogeneity and Optimal Policy	1031
52.2.1	Spatial Heterogeneity in Land Conservation	1031
52.2.2	Effect of Space on Market-Based Solutions	1033
52.3	Spatial Elements of Nonmarket Valuation	1034
52.3.1	Hedonic Valuation	1034
52.3.2	Travel-Cost Analysis	1036
52.3.3	Stated Preference Valuation Techniques	1037
52.4	Spatial Empirical Identification Strategies	1038
52.4.1	Environment and Health	1039
52.4.2	Evaluations of Protected Areas and Payment for Environmental Services Programs	1040
52.5	Models of Behavior in Space	1041
52.5.1	Spatial Sorting Models	1042
52.5.2	Behavior in Land Use and Conservation	1043
52.6	Conclusions	1045
52.7	Cross-References	1046
	References	1046

Abstract

Environmental and natural resource economics has long wrestled with spatial elements of human behavior, biophysical systems, and policy design. The treatment of space by academic environmental economists has evolved in important ways over time, moving from simple distance measures to more

A.W. Ando (✉) • K. Baylis

Department of Agricultural and Consumer Economics, University of Illinois at Urbana-Champaign, Champaign, IL, USA
e-mail: amyando@illinois.edu

complex models of spatial processes. This chapter presents knowledge developed in several areas of research in spatial environmental and natural resource economics. First, it discusses the role played by spatial heterogeneity in designing optimal land conservation policies and efficient incentive policies to control pollution. Second, it describes the role space plays in nonmarket valuation techniques, especially the hedonic and travel cost approaches which inherently use space as a means to identify values of nonmarket goods. Third, it explains a set of quasi- or natural-experimental empirical methods which use spatial shocks to estimate the effects of pollution or environmental policy on a wide range of outcomes such as human health, employment, firm location decisions, and deforestation. Finally, it describes spatial models of human behavior including locational sorting and the interaction of multiple agents in a land use/conservation setting. The chapter ends with a discussion of some promising future areas for further evolution of the modeling of space in environmental economics.

52.1 Introduction

Space is a key dimension of the physical, ecological, and human processes that affect environmental quality and the health of natural resource stocks. Therefore, environmental and natural resource economics has long wrestled with spatial elements of human behavior, biophysical systems, and policy design. The treatment of space by academic environmental economists has evolved in important ways over time, moving from simple distance measures to more complex models of spatial processes.

Researchers have long recognized that the environment is connected to space. Whether because of the distribution of resource quality across space, differential pollution loads, or site-specific policies, space and location matter in environmental and resource economics. Further, there are spillover effects across space; emissions from one place can affect environmental quality in neighboring locations, and fragmentation can degrade the habitat benefits of a given area of conserved land.

Spatial work in environmental and natural resource economics has evolved over time. To take space into account, theoretical work by environmental economists began by including simple spatial resource heterogeneity and contiguity in research on optimal policy design. Initially, heterogeneity was defined as a simple uniform distribution over space, and a single contiguous area was assumed to generate higher ecosystem or habitat benefits than fractured parcels regardless of proximity or the intensity of intervening land uses.

Much of the early empirical work that used space came in the form of hedonic regressions to value location-specific environmental amenities. As a first step, as with the theoretical work, space was usually defined in terms of distance from environmental features or location in certain polygons of the landscape. Spatial empirical work advanced with the introduction of spatial econometrics.

Many empirical papers in environmental economics began to take space into account, initially treating it as a nuisance parameter that generated spatially correlated error terms (Anselin 2002) instead of as an informational component of the data generating process.

Later innovation in environmental economics adopted more nuanced and detailed treatments of spatial processes. For example, research began to differentiate between neighbors on the basis of the direction of pollution flows. Detailed modeling of the spatial nature of ecosystem services, such as habitat provision, is also becoming more common in the literature. Thus, instead of simply controlling for spatial interactions based on a predetermined definition of “neighbors,” authors are now justifying why and how space might affect their model or empirical results, drawing from relevant literatures on natural processes or human interactions.

The most recent step in the evolution of spatial environmental and natural resource economics is the identification and estimation of strategic behavior over space. The idea that location affects land use has been around since von Thuenen. Recent work allows for human migration in response to transportation costs or differential preferences. For example, we have seen large growth in research on locational sorting. Another literature has begun to explore spatial strategic behavior in the subfield studying land use. Some work addresses how actors respond to land use changes or policies and incorporates those reactions into models that target land for conservation. Recent papers have also begun to take the existence of multiple policy makers, private agents, and possible strategic responses into account to better reflect the multitude of principals and agents that collectively affect land use decisions.

In this chapter, we present knowledge developed in several areas of research in spatial environmental and natural resource economics, emphasizing areas that have been and continue to be foci of active research in recent years. We begin with models of simple spatial heterogeneity, starting with a discussion of optimal land conservation policies and moving to analyze how special heterogeneity affects efficient pollution trading. We next discuss the use of space in nonmarket valuation techniques, especially the hedonic and travel cost approaches which inherently use space as a means to identify values of nonmarket goods. The third section of the chapter explains a set of quasi- or natural-experimental empirical methods which use spatial shocks to estimate the effects of pollution or environmental policy on a wide range of outcomes such as human health, employment, firm location decisions, and deforestation. Originally common in labor economics, these methods have been increasingly adopted in environmental economics as an alternative, or at times a complement, to the hedonic approach. Finally, we describe spatial models of human behavior including locational sorting and the interaction of multiple agents in a land use/conservation setting. We conclude with a discussion of some promising areas for evolution of the modeling of space in environmental economics. While this chapter is by no means comprehensive, it is intended to give the reader a sense of how space is treated in modern environmental and natural resource economics.

52.2 Spatial Heterogeneity and Optimal Policy

52.2.1 Spatial Heterogeneity in Land Conservation

Early work in environmental and resource economics determined how to choose conservation and reserve sites optimally when costs and environmental benefits are heterogeneous across space. Simple computational optimization routines can be used to choose sites or spatially target conservation funds to generate the maximum environmental benefits possible, often taking account of complementarities between multiple parcels in the landscape. With fixed parameters of the problem – budget size, benefits, and costs of conserving the parcels – optimal site selection routines will select sets of parcels that have high benefit-cost ratios where benefits consider both the quality of ecological resources on a parcel and the likelihood that the parcel would be degraded in the absence of conservation (Polasky 2005).

Analysis of protected-area network design can, however, also account for the role of space in complex ecological processes when such processes have important effects on optimal design. Production of ecosystem services from reserves often depends on the configuration as well as the total area of lands that are protected. Thus, the integer programming models used for optimal reserve-site selection have been enriched to favor patterns of land that display certain levels of agglomeration.

Sophisticated versions of this work use programming models to choose cost-effective terrestrial reserves in light of detailed spatial idiosyncrasies of the conservation target at hand. Such a model can include details about the population dynamics of the species which is the focus of conservation activity, and models of how the species population depends on proximity to certain features of the landscape and the quality of the unprotected land that lies between reserves. Such models should also incorporate information about spatial heterogeneity in economic use and value. The outcome of such an analysis is identification of the network of lands that maximizes economic surplus in the area while satisfying ecological requirements related to species survival (Albers et al. 2010). Research in marine environments can also use spatial patch population dynamics, specifically knowledge of species source/sink features of different areas in a marine landscape to help policy makers design fishing regulations (including marine reserves) that serve to protect overharvested species and improve social surplus in commercial fisheries (Grafton et al. 2005). In both terrestrial and marine analyses, the best policy is not to place spatially homogenous restrictions on human behavior (protect all wetlands, reduce total fishing effort). Instead, heavy protection of core habitat (or population source sites) can be cost-effective approaches to increasing species populations (and possibly sustainable economic harvest rates), though attention must be paid to patterns of species dispersal through space when designing such policies.

Spatial environmental and natural resource economics has also developed tools for optimal non-reserve policy design that account for important spatial phenomena. For example, economic theory helps us understand how to make spatially

explicit conservation payments. Given that the marginal benefits of conservation on one parcel often depend on the spatial configuration of conservation (or lack thereof) on neighboring parcels, voluntary conservation programs can yield patterns of conservation that are suboptimally fragmented. Effective policies can offer payments for conservation activities which depend on the status of nearby lands. Policy makers can offer agglomeration bonuses – extra payments to landowners for conservation if neighboring parcels are also conserved. Such payments provide incentives that can yield less fragmented patterns of conservation in a landscape (Albers et al. 2010); they may also, however, yield uncompetitive behavior in the bidding process. Auction mechanisms have been developed to provide incentives for agglomerated voluntary conservation while encouraging competitive bidding to minimize rent transfer from the conservation agency to the land owners (Reeson et al. 2011).

In addition, policies are sometimes needed to protect natural resources from threats. Economists have studied how to design such policies efficiently when spatial features of the threats are important. For example, developing countries establish parks and protected areas within which extraction of natural resources is illegal, but it can be difficult to design a cost-effective policy to prevent illegal extraction on the part of nearby villagers. Because extraction activities are carried out by people on foot, there is a strong spatial component to the costs and benefits of extraction in different places within a park. Optimal enforcement may be concentrated in a ring excluding the center and the perimeter of the park; for most cases, the commonly used spatially homogeneous enforcement strategy is highly inefficient (Albers et al. 2010). Policies to control threats to natural resources from invasive species should be spatially explicit as well, using information about spatial heterogeneity in the expected costs and benefits of invasive species control to focus on invasive-species detection and control activities cost-effectively (Kaiser and Burnett 2010).

Finally, spatial environmental economics makes clear that we need to be careful about spatial features of some policies designed to reduce pollution. For example, we would expect development of renewable energy sources such as solar and wind farms to reduce air pollution from electricity generation and thus might put policies in place to encourage such investments. However, spatial idiosyncrasies of the national power transmission grid are such that renewable investments in some locations could actually increase total emissions from that sector by changing the intensity by which some existing fossil fuel-powered plants are utilized. Because the marginal benefits to society of renewable energy installations are spatially heterogeneous, incentives for renewable investments should be as well (Blumsack and Xu 2011).

52.2.2 Effect of Space on Market-Based Solutions

A standard result in environmental economics is that when pollution generates negative externalities – costs not borne by the polluter – then inefficiently large amounts of pollution will be produced by an unregulated market. In the simplest of

cases, the problem of negative externalities from pollution can be solved by imposing a tax on pollution equal to the marginal external cost of the pollution evaluated at its efficient level (Cropper and Oates 1992).

However, pollution and resource use often have spatially heterogeneous negative externalities. For example, air pollution is more harmful if it blows directly into populations of people, and water pollution is more harmful if emitted directly upstream from a sensitive receptor like a lake. Under these circumstances, the optimal policy response is not homogenous across space. Optimal pollution taxation in the spatial context, for example, might not only affect the quantity of emissions, but also shift the location of those emissions. Suppose the harm done by pollution increases with proximity of the emitter to an urban area. If emissions closer to the city have a higher marginal damage, they should be taxed at a higher rate. The difference in taxes would effectively flatten the slope of the transportation costs to the urban center, altering the standard von Thuenen rings of economic activity around the center (Geoghegan and Gray 2005). However, this approach takes the location of the urban center as given. Tax policy could also alter the location of people who are affected by pollution; in some cases, it is more efficient for affected persons to relocate than for the sources of pollution to be moved.

Market-based solutions to externalities such as creating tradable pollution permits are an alternative to taxation. Like the design of optimal taxation, market-based approaches to environmental regulation are complicated by heterogeneous spatial effects of pollution. While a simple trading regime would allow one polluter to buy a permit for 1 unit of emissions from another firm who reduces emissions by 1 unit, if these firms are in separate locations and the effect of emissions is not homogenous across space, this simple trading regime will not result in the optimal distribution of pollution among sources. For example, it is clearly not optimal to trade off 1 unit of emissions in a low-impact area against 1 unit of emissions in a region where pollution causes more harm. Thus, efficient trading can be complicated for pollutants that have specific regional impacts. One policy solution is to divide an area into subregions and only allow trading between sources that are in the same region, but this approach has the potential cost of creating thin markets. Another approach is to insist that pairs of sources trade permits at ratios that accurately reflect heterogeneity of marginal damages caused by pollution from different sources, but this solution creates administrative complexity. Spatial heterogeneity presents policy-makers with trade-offs: charging firms their true marginal damage yields efficiency gains, while increasing the costs of complexity (including the need for increased monitoring), and raising concerns about distributional features of spatially heterogeneous policies (Olmstead 2010).

52.3 Spatial Elements of Nonmarket Valuation

Even before spatial analysis gained prominence in economics, some nonmarket valuation techniques (such as hedonic analysis and the travel-cost method) were intrinsically spatial. Environmental economists have enhanced the use of space in those methods over time, and spatial concerns have been incorporated into other

nonmarket valuation tools as well (Bateman et al. 2006). This effort has been facilitated by the development of a wide range of tools for applied spatial data analysis and econometric regression (Fischer and Getis 2010).

52.3.1 Hedonic Valuation

Hedonic housing price analysis is grounded in the economic intuition that the price of a house will be a function of all its features including the environmental quality and access to natural amenities that are associated with its specific location in space. Sellers choose features to supply to maximize profit; buyers choose which house to buy (for a given price) to maximize utility. The market equilibrium yields a hedonic price function (price as a function of attributes) that can be estimated econometrically using spatially explicit data on houses, their sales prices, their conventional attributes (e.g., number of rooms, square footage), and their environmental attributes. One can interpret the marginal effect on price of an environmental feature as the marginal willingness to pay of people in this market for that feature. These marginal willingness-to-pay measures inform us about the welfare effects of highly localized changes in the environmental quality. However, it is notoriously difficult to use hedonic analysis to estimate the welfare effects of a widespread change in environmental conditions (e.g., cleaner air in all of Southern California) because the market equilibrium would change and create an entirely new hedonic price function which can be difficult to predict from current conditions (Cropper and Oates 1992; Palmquist 2005).

Observations in hedonic analyses can display spatial autocorrelation because of two processes. A spatial lag process arises when the outcome observed in one location is a function of the outcome of neighboring locations. For example, the price of one house may directly affect the price of the neighboring houses, perhaps by updating seller information about current market values. A spatial lag can also arise through the common use of a resource, such as neighbors competing with each other in the use of irrigation water (Anselin 2002). In contrast, a spatial error process refers to spatial correlation in the residuals. In the hedonic analysis literature, several studies have used econometric approaches that take into account possible spatial autocorrelation from both sources. Failure to account for autocorrelation can yield inconsistent estimates of the coefficients on environmental quality, while failure to capture spatial lag lead to bias, meaning that the estimates of the marginal effects of changing environmental quality in one location are missing spillovers into neighboring properties (Anselin 2002).

Estimating how much pollution affects a specific house is nontrivial, since most pollution is usually only measured at a few locations in space. Thus, pollution measures are often spatially interpolated from these point data using kriging to generate an estimate of pollution at any specific latitude and longitude. Another approach to dealing with limited pollution data is to analyze housing prices within a larger spatial unit that either conform more closely to the point data or use geographic averages of the point measures.

One concern is that, like all interpolated variables, these environmental variables are measured with error, and this error may well be correlated with other unobservables that are also correlated with housing prices. For example, houses on a certain ridge could be subject to cooling ocean breezes that also result in a highly localized drop in pollution. The potential heteroscedasticity induced by using estimates for pollution can be addressed by correcting for both spatial and heteroscedastic error terms. However, the more fundamental concern about omitted variable bias remains. Such bias may be present even without interpolated environmental variables, for there may always be important location-specific unobserved variables that are correlated across space with both the housing price and the environmental characteristic. The problem of omitted variables can be addressed by using repeated sales of the same house over time, or by including other regional fixed effects.

Traditional hedonic analysis has employed fairly simple notions of location, space, and neighbors. For example, it has usually used measures of environmental quality onsite (e.g., air pollution levels) or simple distance to an environmental amenity or disamenity (e.g., open space, hazardous waste site). However, such simple definitions may fail to capture important effects. For example, the walking or driving time to a park might affect the price of a house more than the Euclidian distance, and having an amenity across a major road might increase the perceived distance of that amenity more than having it across a minor street. While the value of water quality improvements in a lake is diminishing with the distance of a house from a lake, there may be a discontinuous jump in value at the waterfront; there is often a complex story to be told about the actual ecosystem services that are being valued through the proxy of pollution measures (recreation, visual aesthetics, ecological health) and the role that space plays in mediating people's experiences of those services. Furthermore, when estimating how house prices might affect each other, such as when estimating a spatial lag, houses on the same block might affect each other's values more than houses one block over even if they are the same distance apart.

Such concerns can be addressed by taking a broader spatial view of the ways in which environmental quality might affect the relative desirability of homes in a housing market, and by taking care to define variables in hedonic models to reflect spatial realities and processes on the ground. The effects of pollution may not be simple – neither uniform, nor merely a matter of being in a polygon that is contiguous with a source, nor a linear function of distance from a source. In such cases, one can use detailed information on the dispersion of the effects of pollution to inform a hedonic analysis that estimates people's willingness to pay to reduce it.

The hedonic spatial model can also be enriched by enhancing the interaction between space and time, extending the standard hedonic model to allow households to be forward looking and to face transaction costs of moving (Bishop and Murphy 2011). Under such plausible circumstances, households weigh the cost of an environmental amenity (captured by the price premium associated with houses in locations with good environmental values) against the present discounted value of

the stream of future utility they will obtain from the amenity. Incorporating forward-looking behavior yields much bigger estimates of consumer marginal willingness to pay for a spatially heterogeneous environmental amenity.

52.3.2 Travel-Cost Analysis

The other nonmarket valuation that is most intrinsically spatial is the travel cost approach to estimating the values people place on the quality of natural resources. This method estimates demand for recreational sites such as beaches, lakes, and forests as a function of features of those natural sites; the results yield estimates of the values of the features (e.g., water quality, species populations) included in the analysis. The travel-cost approach uses data on how often people visit the sites of interest and how much those visits cost each individual in the data set, where travel cost depends in part on how close someone lives to a site. Single-site models use econometric analysis to estimate how quantity of visits to a site depend on environmental quality; multiple-site models use a random-utility model (RUM) econometric approach to estimate how the choice of which of several sites to visit depends on the attributes of all the sites and how much travel to them costs (Cropper and Oates 1992).

Travel-cost valuation methodology has evolved to include new features of space. The cost of travel was always measured as a function of how far a person lives from a site, but if people engage in locational sorting, distance from a site (and hence measured travel cost) will be correlated with unobservable preference heterogeneity, creating biased coefficient estimates. Latent class models can be used to control for this endogeneity (Barenklau 2010). Other problems can arise if multiple sites between which people choose for recreation (e.g., patches of a forest for hunting, lakes in a chain for fishing) are connected physically and ecologically across space. If, for example, a change in water quality at one lake causes fish populations to change and redistribute through an entire chain of lakes, then conventional travel-cost analysis can yield misleading information about the welfare effects of that change. A structural model of recreation site choice and harvest intensity must be coupled with a spatial model of population dynamics to understand the welfare effects of making improvements to features of one or more sites in such a system (Albers et al. 2010).

52.3.3 Stated Preference Valuation Techniques

Stated preference valuation methodologies (contingent valuation and choice experiment studies) use information from hypothetical survey questions to estimate consumers' willingness to pay for environmental goods and services even if the values they gain are not based in any way on direct use (Cropper and Oates 1992). Nonuse values may not be affected by distance to environmental amenities. However, distance may play a factor in the values people place on the environment if

people have a localized “sense of place” or if use values comprise a large fraction of the total value people place on environmental public goods.

Thus, space is recognized now to be an important part of even stated-preference valuation approaches. Data on how far people are from the amenities to be valued can be included directly in the specifications of such studies to measure how distance affects environmental goods and ascertain how that effect varies with income. Including distance explicitly in individual willingness-to-pay functions helps cost-benefit analysts avoid making arbitrary choices about the spatial extent of the population of people that are affected by a project (Bateman et al. 2006). The value people place on an environmental improvement may also depend on spatial variation in the current quality they experience for the amenity in question.

52.4 Spatial Empirical Identification Strategies

As noted in Sect. 52.3.1 on hedonic analysis, space or location has long been used as a source of information to identify and estimate the effects of variation in environmental quality. As an alternative to the more structural hedonic model, the last decade has produced substantial growth in the application of quasi or natural experiments to estimate the effects of pollution and environmental policy. Spatial variation can be used to identify the effect of a treatment such as a policy shift or change in environmental conditions (Smith 2007). If policies or shocks are specific to a location, it is possible to compare outcomes in these areas to outcomes in other untreated locations to measure the effect of the treatment. If the outcomes are observable before and after the treatment, one can control for time-invariant observables which can often confound estimates obtained from other approaches. Standard policy evaluation procedures (such as difference-in-difference, matching, or regression discontinuity methods) can then be applied to estimate the effect of the treatment.

Matching is a technique that compares treated with control observations on the basis of their observable characteristics. This technique addresses potential bias that might arise due to systematic differences in covariates between the treated and control observations. It does not, however, address the concern that treatment might be related to some unobservable characteristic that in turn affects the outcome of interest. A difference-in-difference approach compares treatment and control observations before and after the introduction of the treatment. This approach controls for time-invariant differences between the treated and control observations. Regression discontinuity design makes use of a fixed threshold that determines whether an observation is “treated” or not. For example, if the treatment occurs when an individual turns 65, one can use the outcomes of 64½-year-olds as controls. For a discussion of these and other program evaluation techniques, see Khandker et al. (2010).

The shocks used for identification in spatial environmental and natural resource economics have ranged from a decrease in pollution (e.g., from a recession or a localized plant closure), to natural disasters, and to the introduction of protected

areas. Along with measuring the effect of policies on intended outcomes, the use of this quasi-experimental technique has been applied to estimate nonmarket valuation of environmental amenities and health outcomes. By definition, one requirement of the quasi-experimental approach is that when using variation across space as a source of variation, one needs a spatially varied shock. For example, spatially heterogeneous policies, such as air-pollution emission standards that vary non-attainment status of a country, have become popular sources of identification to estimate willingness to pay for pollution or the influence of pollution on health or economic activity.

While it has some advantages, the quasi-experimental methodology has limitations as well. One challenge is in choosing the appropriate spatial scale for analysis. Often researchers cannot observe responses at the individual level and use regional values instead. At least two problems arise from this. First, patterns of correlation among variables across space are not always robust to the spatial units over which the data are aggregated. This problem of ecological fallacy (Anselin 2002) is most pronounced if individual variation within a region is large compared to the variation among regions. Second, non-parcel level data may not be fine enough to observe the effects of some environmental shocks (Smith 2007). Quasi-experimental studies may also yield biased results if they assume treatment effects that are constant with distance when, in fact, both the treatment itself and the impact of a treatment on housing prices are idiosyncratic across space (Auffhammer et al. 2009; Smith 2007).

Last, one crucial assumption required for the use of quasi-experimental methods is that the treatment is not assigned based on unobservables that also affect the outcome. While some random shocks, such as weather variation, may well fall into this category, other shocks (such as a regional policy, the shut-down of a plant, or spatially delimited critical habitat for endangered species) are potentially more problematic. If those unobservables are time-invariant, the use of fixed effects may mediate the problem. Fixed effects, however, do not solve the problem of unobserved variation generating a differential effect of observed characteristics on the outcome. For example, if unobserved political influence affects the location of a new environmental policy and political influence also determines how that policy affects economic outcomes, one could still estimate a biased coefficient for the effect of the policy on economic outcomes even with fixed effects.

52.4.1 Environment and Health

Arguably the largest growth area in the use of these natural or quasi-experiments in environmental economics has been on measuring the effect of pollution on health. As with the willingness-to-pay literature, this is a topic that has previously seen the broad application of hedonic analysis. There is a substantial literature that measures the costs of environmental health risks and disease that use epidemiological methods to estimate a dose-response function of, say, exposure to a chemical and health outcomes, and then use wage hedonics to estimate the perceived costs of

those work-related risks (Viscusi and Gayer 2005). Other papers have estimated the cost of health effects using variation in housing prices.

Various authors have used natural experiments arising from a temporary plant closure or changes in traffic patterns to estimate the effect of emissions on health outcomes. Other authors have used economic downturns as an instrument for changes in county-level pollution to estimate the effect of pollution on health. As with the other quasi-experimental studies, one concern is finding the appropriate scale of analysis. More recent papers make use of smaller scale variation in pollution levels, using within-zip-code or school district variation to be better able to control for other neighborhood fixed effects (for example, see Currie et al. 2009).

Another approach is to use natural and environmental disasters as a source of variation to estimate the effect of these disasters on health outcomes. A continuing challenge is how exactly to model the spatial and temporal exposure to these shocks and to address human responses to either the threat or incidence of exposure (such as migration). In using this methodology, researchers also need to be careful to rule out potential spillovers resulting from the treatment into neighboring control regions; such spillovers could render the control group uncontrolled, and therefore bias the estimate of treatment effect.

52.4.2 Evaluations of Protected Areas and Payment for Environmental Services Programs

Spatial analysis has been and can be used to estimate the effectiveness of conservation measures in preventing environmental degradation such as deforestation. The methodology has been developed to study programs that establish protected area policies that offer payments to landowners for activities that preserve or increase flows of environmental services – “payment for environmental services” (or PES) programs. Location-specific attributes and the spatial process of land use play important roles in estimating the effects of these programs.

Early evaluations of conservation efforts compared outcomes (such as deforestation rates) in areas subject to a conservation measure, such as legal protection, to outcomes in plots outside the boundaries of this protection. The problem with this approach is that protected and unprotected areas frequently differ in ways that systematically bias the comparisons (Andam et al. 2008). For example, countries may naturally place their protected areas in regions that face lower deforestation pressure (Joppa and Pfaff 2010). In these circumstances, estimates from a simple comparison of outcomes inside and outside of the protected area boundaries would overstate the impact of conservation policies. To overcome these biases and develop more accurate comparisons, conservation research must consider realistic counterfactual scenarios (Ferraro 2009). Thus, researchers must adopt evaluation techniques that permit comparison of observed outcomes with what would have happened in the absence of a conservation effort. The difficulty lies in that counterfactuals cannot be observed directly and instead should be carefully estimated.

Recent research has attempted to estimate a counterfactual in evaluations of conservation programs. Costa Rica's payment for environmental services program has been assessed using linear regression models and two types of matching estimators to compare the deforestation rates of communities that participate in the program and communities that do not, controlling for observable features of the landscape such as slope, distance to cities, and ecological zones; the results indicate the program had little effect on deforestation (Andam et al. 2008). It is also possible to take an explicitly spatial approach to the analysis of conservation program effectiveness. One technique is to control for possibility of spatially autocorrelated errors in the regressions that analyze the impact of conservation policy on landscape degradation (Alix-Garcia 2007). A second approach is to control for spatial spillovers from one observation to the next, by explicitly estimating the spatial lag associated with land use change. Failure to control for such spillovers has been found to have large effects on the estimates of treatment effects (Honey-Roses et al. 2011). A third spatial strategy is to estimate the effect of the program on nearby areas, or explicitly estimate the leakage caused by the policy. If there is a spatial lag process associated with deforestation, land use in observations on the boundary of the treatment area might well be affected by the treatment of the neighboring area, implying that they are not appropriate control observations.

52.5 Models of Behavior in Space

Until now, this chapter has largely focused on models where spatial effects arise from features of nature. Such models assume that resource locations are given and that the heterogeneous effects of pollution are determined by factors exogenous to humans, like wind or hydrology. However, spatial heterogeneity may arise from human behavior and the resulting economic forces. Research in environmental and natural resource economics has developed understanding of various spatial dimensions of human behavior.

From the simplest von Thuenen model of land use being driven by variation in transport costs to market to the rise of New Economic Geography in the 1990s, we now have models that predict the growth of cities. The New Economic Geography approach models population centers as arising from tension between agglomeration economies (driven by monopolistic competitive firms) and congestion costs. These models still assume at their base a featureless plain, where migration is driven by differences in real wages. Once one introduces an influential spatial feature, people with a strong preference for that amenity may migrate for other reasons. This innovation has led to the concept of spatial sorting.

Economics predicts that people respond to incentives. Incentives may themselves arise from features of the landscape other than just proximity to the nearest urban center. For example, zoning and other land-use rules may place restrictions on the use of some land, pushing these land uses elsewhere (an effect also known as leakage). As some land is removed from potential development, the price of

development rights may increase in other regions. These and other behavioral responses are incorporated into modern models of land use and land conservation research.

Regulation of environment and natural resource use is complicated by the existence of multiple regulators and multiple regulated actors, giving rise to the potential for strategic behavior and collective action problems. These problems gain an extra dimension of complexity when the cooperation or competition occurs over space. Spatial environmental and natural resource economics now incorporates some of these multiagent behaviors in space.

52.5.1 Spatial Sorting Models

One recent thread of research in the field of environmental and natural resource economics has rapidly become an established and influential feature of the literature: spatial sorting models (Palmquist 2005). This body of work evolved from early work by Tiebout (Banzhaf and Walsh 2008) on how people “vote with their feet” and move to places that have bundles of attributes – including environmental quality and cost – they prefer. Modern spatial sorting models are theoretically and computationally complex, and are used for a wide range of functions.

One category of research on sorting models is positive – just seeking to describe whether (and if so, how) people sort across space in the face of spatial heterogeneity of attributes. This research can help us to understand the forces that drive demographic patterns within urban areas, and shed important light on questions of environmental justice. These models can also be used to explore how proposed changes in environmental quality will affect the distribution of people in the landscape and their subsequent well-being.

Early theoretical models of spatial sorting equilibria assumed that households have heterogeneous incomes and preferences over housing and public good characteristics of a location. Communities vary in how expensive they are and in the level of the public good they provide. Individuals choose where to live to maximize their utilities subject to their budget constraints; housing prices in communities adjust until equilibrium is reached such that no household would prefer to live somewhere other than where they are living. Even in the simplest models, assumptions must be made about the structure of indirect utility functions in order to ensure that equilibrium exists. The models also assume (implicitly or explicitly) that all households have perfect information about community characteristics and the preferences of other households, all households are able to purchase as much housing as they want in their preferred locations, and moving is costless. The resulting equilibria have communities that are stratified by income if preferences are homogeneous, and households sorted differentially according to the features they care most about if preferences vary (Palmquist 2005).

Later models (e.g., Bayer and Timmins 2005) allow for spillovers between individuals that choose a given location; spillovers can either be positive

(as in the case of agglomeration economies) or negative (if there is congestion). Under these circumstances, multiple equilibria are often possible, particularly if there is a strong agglomeration effect. One can still use data to estimate the features of models that have multiple equilibria, but multiplicity makes it more difficult to draw conclusions about what the re-sorting effects will be of major changes in a region such as cleaning up a hazardous waste site.

Empirical work has sought to identify whether sorting behavior in response to spatial environmental heterogeneity is an important factor in residential markets. Econometric approaches to this problem include statistical analysis of changes over time in socio-demographic and housing characteristics of locations near sites that experience changes in environmental quality (Banzhaf and Walsh 2008). There is evidence that people locate at least partly in response to environmental features of neighborhoods, and that such dynamics can exacerbate income segregation in urban areas.

Because of their utility-theoretic underpinnings, sorting models have been used as the foundation for a new approach to estimating the values people place on elements of environmental quality that do not have market values. Researchers can use neighborhood-level land value data to obtain structural estimates of the parameters underpinning residential sorting models and thus estimate values of spatially differentiated environmental amenities such as air quality and open space (Klaiber and Phaneuf 2010). In addition to generating value estimates that can be used in cost-benefit analyses, this research reveals several insights about environmental policy and research. First, the benefits of an environmental improvement policy depend on how it is distributed in space. Second, benefit estimates based on traditional nonmarket-valuation techniques may be incorrect if the environmental changes to be valued are large enough to induce significant re-sorting. An example can illustrate. Suppose air quality in the neighborhood of Gryffind is originally much lower than in Slyther; people would sort such that the people who value clean air most intensely would pay a premium and live (disproportionately) in Slyther. If we improve air quality in Gryffind, there is initially just a small welfare increase because the people who live there care relatively little about air pollution. With resorting, there are two effects: (1) The people who value air quality more highly move to Gryffind, and thus the benefit to residents there is higher. (2) Housing prices fall in Slyther and rise in Gryffind, causing indirect price effects on welfare that depend in size and spatial distribution on details of the situation.

The structural sorting-equilibrium approach does have the great advantage of taking dynamic factors into consideration. However, it requires analysts to impose much structure on the underlying model and to make arbitrary choices about the boundaries over which communities (which are the unit of observation) are defined. This latter activity may be extremely problematic given that results of spatial statistical analysis have long been known to be sensitive to the manner in which data are aggregated across space (Anselin 2002). Future work on this methodology may seek to resolve these issues.

52.5.2 Behavior in Land Use and Conservation

Land use is an area that straddles several disciplines in economics (urban economics, environmental economics, and economic geography) and has long recognized the importance of human interaction with space. Early models of land use often ignored the behavioral component and were largely meant to fit, as opposed to explain, the data.

More recently, models of optimal conservation planning have been developed that incorporate spatial heterogeneity of environmental costs and benefits with spatial economic models of the probability of land use change. Instead of merely conserving land based on selecting parcels with the highest environmental benefit per dollar, it improves economic efficiency to target those parcels with the highest environmental benefit per dollar that are also under the highest threat of development.

Other models of land-use change have begun to take into account behavioral responses to development or development policy changes. In general, restrictions on land use in one part of space (such as zoning) can intensify the limited activity in other areas that are not controlled by the restrictions; this is the generalized phenomenon of leakage. Some land-use restrictions, such as urban policies mandating embedded open space, can increase the value of development in neighboring areas so much that they accelerate leapfrog urban sprawl (Irwin et al. 2009).

Finally, some research incorporates the fact that multiple actors are involved in conservation, and that these actors likely interact, and often interact strategically. Spatial strategic behavior is best known in models of how local governments set their levels of public goods, taxes, and/or regulatory stringency. If firm location choice is endogenous, nearby jurisdictions may compete on the level of taxation and public goods. This competition is further complicated by economic activity induced by firm location having spillovers to neighboring locations. Strategic private responses can thwart governments in many of the actions they try to take to improve environmental quality, creating hold-up problems when an agent is trying to establish an agglomerated protected area that requires buy-in from multiple land owners, sometimes shifting private conservation into parts of a landscape that are spatially disparate from the locations of public conservation activities (Albers et al. 2010).

The most recent generation of research on conservation reserve design uses economic theory to inform the strategic choice of lands for reserves taking into account the spatial responses of multiple human agents to those choices. Empirical research has identified many ways in which human behavior in space responds to changes in the environment; for example, the establishment of government-protected lands can increase the price of land and the threat of development (or likelihood of conservation) in the area (Irwin et al. 2009). Thus, optimal reserve choices by one agent should be strategic, taking into account the likely responses of other agents (Albers et al. 2008) and likely changes in the land market which affect the risk to other parcels of conversion and the cost to the decision maker of future conservation (Armsworth et al. 2006). Such strategic decision making can yield improved conservation outcomes, but can entail making seemingly counterintuitive choices such as avoiding putting protected areas in some locations with high

ecological value. Similarly, econometric work has documented how harvesting activity varies across space with changes in factors such as target (e.g., fish) populations and the presence of regulations such as spatial closures (Grafton et al. 2005; Albers et al. 2010). Endogenous harvesting behavior affects the outcomes of spatially explicit harvesting regulations – if one area is closed, harvesters work more intensively in another area, and if regulations increase target populations, harvesting effort will increase. Socially optimal spatial resource use regulations can be designed in ways that take such endogenous behavior into account (Grafton et al. 2005).

52.6 Conclusions

Some areas for future work in spatial environmental and natural resource economics seem to be particularly important and promising. In the area of spatial policy evaluation, one area where future work is needed is to more formally incorporate spatial data-generating processes into the quasi-experimental setting. For example, the use of propensity-score matching (PSM) is potentially biased in the presence of spatially correlated error terms. Just as a probit estimation generates potentially biased estimates in the presence of heteroscedasticity, the initial probit regression used to generate the propensity of treatment may be inherently biased by the presence of spatial correlation. More fundamentally, in the presence of a spatial lag, estimates will likely be biased, and further, control observations neighboring treated regions may themselves be affected by the treatment (Honey-Roses et al. 2011). The bias in this instance could go either way depending on the nature of the lag process. While this spatial effect may complicate difference-in-difference analyses, it is even more potentially problematic for regression discontinuity design where the regression discontinuity is spatial in nature. Note that since the amount of spillover is not constant over time, and may be directly affected by the treatment, using observation-level fixed effects does not solve the bias.

A related area of concern is that a treatment itself may actually change the scale and scope of important spatial processes related to that treatment. For example, a fuel tax may affect the degree of spatial spillover from economic activity in one area to economic activity in neighboring areas by changing patterns of commuting behavior. These effects on spillovers may be substantial and have large effects on policy outcomes which have not systematically been studied.

In the area of spatial policy design, truly optimal policies need to take spatial strategic reactions into account rather than treating other actors as merely reactive. Papers that apply game theory to spatial policy decisions are rare (Albers et al. 2008); more work needs to be done in this area. For example, private actors are known anecdotally to buy land for speculation if they anticipate conservation agents wanting to buy it for protected areas. This phenomenon is different from that of markets responding to conservation with increased prices nearby and should be worked into spatial-dynamic models of optimal reserve-site selection.

Future work in spatial environmental and natural resource economics may even move to redefine what we mean by “space.” Extant research and knowledge in this field conceptualizes space in traditional geographic terms. However, other dimensions of space exist that may affect natural processes and human behavior. Economic interactions may facilitate technological adoption more than mere geographic proximity. Social distance and social networks can affect attitudes and behavior through facilitating both information flow and influence. As an example, information and influence can affect individual’s valuation of a disamenity such as hazardous waste. Further, social influence can be used to improve monitoring, enforcement, and, therefore, management of a local common pool resource, such as a community pasture. Current research in spatial econometrics is moving forward to allow researchers to estimate spatial weights or spatial spillover patterns, as opposed to merely estimating the degree of spillover given an assumed structure of the extent to which different spatial units function as neighbors. These advances in spatial econometrics will facilitate future research that quantifies the effects of spillovers in environmental and natural resource economics.

Knowledge in spatial environmental and natural resource economics already includes theoretical and empirical models that inform spatial environmental policy design, evaluate policy effectiveness, help us predict human behavior in a landscape, and help place values on environmental goods that are spatially heterogeneous and convey benefits in ways that vary with spatial processes. However, work in this field of research is still very much ongoing and the field is still evolving; much more needs to be done.

52.7 Cross-References

- ▶ Classical Contributions: Von Thünen, Weber, Christaller, Lösch
- ▶ Dynamic and Stochastic Analysis of Environmental and Natural Resources
- ▶ Economic Valuation: Concepts and Empirical Methods
- ▶ Housing Choice, Residential Mobility, and Hedonic Approaches
- ▶ Interpreting Spatial Econometric Models
- ▶ Scale, Aggregation, and the Modifiable Areal Unit Problem
- ▶ The Hedonic Method for Valuing Environmental Policies and Quality

Acknowledgments This chapter is based in part on work supported by USDA-NIFA Hatch project number #ILLU-470-316. Lead authorship is equally shared by the two coauthors. The authors thank the editors of this volume for useful comments on the manuscript.

References

- Albers HJ, Ando AW, Batz M (2008) Patterns of multi-agent land conservation: crowding in/out, agglomeration, and policy. *Resour Energy Econ* 30(4):492–508
- Albers HJ, Ando AW, Shogren JF (2010) Introduction to spatial natural resource and environmental economics. *Resour Energy Econ* 32(2):93–97

- Alix-Garcia J (2007) A spatial analysis of common property deforestation. *J Environ Econ Manag* 53(2):141–157
- Andam K, Ferraro PJ, Pfaff A, Sanchez-Azofeifa AG, Robalino JA (2008) Measuring the effectiveness of protected area networks in reducing deforestation. *Proc Natl Acad Sci* 105(42):16089–16094
- Anselin L (2002) Under the hood: issues in the specification and interpretation of spatial regression models. *Agric Econ* 27(3):247–267
- Armsworth PR, Daily GC, Kareiva P, Sanchirico JN (2006) Land market feedbacks can undermine biodiversity conservation. *Proc Natl Acad Sci* 103(14):5403–5408
- Auffhammer M, Bento AM, Lowe SE (2009) Measuring the effects of the Clean Air Act Amendments on ambient concentrations: the critical importance of a spatially disaggregated analysis. *J Environ Econ Manag* 58(1):15–26
- Banzhaf HS, Walsh RP (2008) Do people vote with their feet? An empirical test of Tiebout's mechanism. *Am Econ Rev* 98(3):843–863
- Barenkraut KA (2010) A latent class approach to modeling endogenous spatial sorting in zonal recreation demand models. *Land Econ* 86(4):800–816
- Bateman I, Yang W, Boxall P (2006) Geographical information systems (GIS) and spatial analysis in resource and environmental economics. In: Tietenberg T, Folmer H (eds) *The international yearbook of environmental and resource economics 2006/2007: a survey of current issues*. Edward Elgar, Northampton, pp 43–92
- Bayer P, Timmins C (2005) On the equilibrium properties of locational sorting models. *J Urban Econ* 57(3):462–477
- Bishop KC, Murphy AD (2011) Estimating the willingness to pay to avoid violent crime: a dynamic approach. *Am Econ Rev Pap Proc* 101(3):625–629
- Blumsack S, Xu J (2011) Spatial variation of emissions impacts due to renewable energy siting decisions in the Western U.S. under high-renewable penetration scenarios. *Energy Policy* 39(11):6962–6971
- Cropper ML, Oates WE (1992) Environmental economics: a survey. *J Econ Lit* 30(2):675–740
- Currie JE, Hanushek A, Kahn EM, Neidell M, Rivkin SG (2009) Does pollution increase school absences. *Rev Econ Stat* 91(4):682–694
- Ferraro PJ (2009) Counterfactual thinking and impact evaluation in environmental policy. *New Dir Eval* 2009(122):75–84. doi:10.1002/ev.297
- Fischer MM, Getis A (eds) (2010) *Handbook of applied spatial analysis: software tools, methods, and applications*. Springer, Berlin
- Geoghegan J, Gray W (2005) Spatial environmental policy. In: Tietenberg T, Folmer H (eds) *The international yearbook of environmental and resource economics 2005/2006: a survey of current issues*. Edward Elgar, Northampton, pp 52–96
- Grafton RQ, Kompas T, Schneider V (2005) The bioeconomics of marine reserves: a selected review with policy implications. *J Bioecon* 7(2):161–178
- Honey-Roses J, Baylis K, Ramirez I (2011) Do our conservation programs work? A spatially explicit measure of avoided deforestation. *Conserv Biol* 25(5):1032–1043
- Irwin EG, Bell KP, Bockstael NE, Newburn DA, Partridge MD, Wu J (2009) The economics of urban-rural space. *Annu Rev Resour Econ* 1:435–459
- Joppa L, Pfaff A (2010) Reassessing the forest impacts of protection: the challenge of nonrandom location and a corrective method. *Ann N Y Acad Sci* 1185:135–149
- Kaiser BA, Burnett KM (2010) Spatial economic analysis of early detection and rapid response strategies for an invasive species. *Resour Energy Econ* 32(4):566–585
- Khandker SJ, Koolwal GB, Samand HA (2010) *Handbook on impact evaluation: quantitative methods and practices*. World Bank, Washington, DC
- Klaiber HA, Phaneuf DJ (2010) Valuing open space in a residential sorting model of the Twin Cities. *J Environ Econ Manag* 60(2):57–77
- Olmstead S (2010) The economics of water quality. *Rev Environ Econ Policy* 4(1):44–62
- Palmquist RB (2005) Property value models, chapter 16. In: Mäler KG, Vincent J (eds) *Handbook of environmental economics*, vol 2. Elsevier, Amsterdam, pp 763–819

- Polasky S (2005) Strategies to conserve biodiversity. In: Tietenberg T, Folmer H (eds) The international yearbook of environmental and resource economics 2005/2006: a survey of current issues. Edward Elgar, Northampton, pp 157–184
- Reeson AF, Rodriguez LC, Whitten SM, Williams K, Nolles K, Windle J, Rolfe J (2011) Adapting auctions for the provision of ecosystem services at the landscape scale. *Ecol Econ* 70(9):1621–1627
- Smith VK (2007) Reflections on the literature. *Rev Environ Econ Policy* 1(2):300–318
- Viscusi K, Gayer T (2005) Quantifying and valuing environmental health risks, chapter 20. In: Mäler KG, Vincent J (eds) *Handbook of environmental economics*, vol 2. Elsevier, Amsterdam, pp 1029–1103

Daria A. Karetnikov and Matthias Ruth

Contents

53.1	Introduction	1049
53.2	Expected Impacts Based on Level of Urbanization	1051
53.2.1	Density-Dependant Impacts	1051
53.2.2	Agriculture and Forests Impact	1053
53.2.3	Natural Landscapes	1054
53.3	Regional Differences in Risk and Mitigation Capacity	1055
53.3.1	North America	1056
53.3.2	Europe	1058
53.3.3	Asia	1059
53.3.4	Latin America	1061
53.3.5	Africa	1062
53.3.6	Australia and New Zealand	1063
53.4	Interconnectivity and Reach of Impacts	1065
53.5	Global Social Justice	1066
53.6	Conclusion	1068
Appendix	1068	
References	1069	

Abstract

The expected global impacts of climate change can be attributed to a set of common stressors. The magnitude of specific impacts, however, depends on the extent to which regional resources – from ecosystems to human-made

D.A. Karetnikov (✉)
University of Maryland, College Park, MD, USA
e-mail: dariakar@yahoo.com

M. Ruth
Department of Civil and Environmental Engineering, School of Public Policy and Urban Affairs,
Northeastern University, College Park, USA
e-mail: m.ruth@neu.edu

infrastructures – are at risk and the abilities of regions to mitigate that risk. This chapter begins with an overview of some of the impacts expected from climate change, stratified by the density of populations and economic activities. Then we review differences in risk and mitigation capacities across major regions. The inherent interconnection of environmental, economic, and social dimensions of climate impacts underscores the need to assess climate change impacts in ways that address these dimensions.

53.1 Introduction

The arguments made by researchers, policymakers, and activists on the need to curb greenhouse gas emissions often revolve around an implicit or explicit understanding of the expected costs and benefits of probable impacts. The result of any structured comparison between costs and benefits, in turn, depends on the impacts that are considered, how these impacts are defined and distributed across the economy, society, and its environment, how they are measured and weighted, and how they are aggregated to generate a *net* cost or benefit that then guides action. On the one hand, many of the expected impacts from climate change can be usefully discussed together – temperature increases, changes in precipitation, and an increase in the number of more severe weather events that threaten to disrupt **urban** and **suburban** centers; **infrastructure networks** like roads, energy transmission lines, and shipping routes; and evolved **ecosystem** relationships important to agriculture and natural landscapes in many densely populated regions. On the other hand, the situational variables, where an impact occurs and whom it affects and to what extent, determine the magnitude of the damages and must be understood in the local context of the capacity to prepare for an event or deal with its aftermath.

Regional variability cannot be subtracted out from global assessments of climate change impacts, and it introduces enormous complexities into the global decision-making process of designing and agreeing on a united response. Each region has its own economic and trade portfolios at risk and its own historical and political value systems to deal with that risk. This chapter presents an overview of some of the impacts expected from climate change that are common across the world based on the level of urbanization – from densely populated urban areas to less-populated regions. Then we review differences in risk and **mitigation capacities** across major regions, how economic and political interconnectivity links regions closer, and the differences in the means that are necessary to mitigate or adapt to the risk and to deal with the impact itself. Much of our analysis relies on existing geophysical models and their estimates.

Many such climate models exist that relate hundreds of complex global parameters together to chart out the circulation patterns between the Earth's atmosphere and its oceanic system and to derive potential responses to the current emission trends. Uncertainty is an inherent part of the modeling process. Because of the differences in how the models treat the various parameters and their relationships, to actually estimate impacts of climate change, **the Intergovernmental Panel on**

Climate Change (IPCC) uses a number of models' projections to arrive at its results. Their details are described by the IPCC itself (Intergovernmental Panel on Climate Change 2007). Beyond the geophysical models, integrated assessment models (like the RICE and the DICE models developed at Yale University, the Stanford-based MERGE model, the FUND model developed by the Dutch economist Richard Tol, or the PAGE2002 model used in the influential Stern review) provide the framework to picture how climatic changes may overlap with societal dimensions (for detailed overviews of such models, see Stanton et al. 2009 and Ortiz and Markandya 2009). Often the models are used to calculate the net or the average impact of climate change – or to perform a benefit-cost analysis, arriving at conclusions that mask regional vulnerabilities.

Among the poignant critiques of such benefit-cost analyses are that the models in the background tend to ignore the existence of low probability but catastrophic outcomes and that the models estimate future damage probability functions relying on normal distribution rather than the so-called fat-tail Pareto distribution that may more accurately represent the uncertainty involved given that the current concentrations of greenhouse gases exceeds any past experience. Simply substituting the type of probability function used alters the estimates for probabilities of temperature changes experienced under different scenarios. For instance, at a concentration of greenhouse gas emissions double the 2011 levels, under the assumption of a normal probability distribution, the probability that global average temperatures will stabilize at 10 °C higher than the 2011 temperature is more than seven times lower than when using a “fat-tail” distribution (Weitzman 2011). Clearly, the size of the temperature increase determines the severity of damages. Other problems with using the models persist, such as finding the proper rate of discounting to estimate long-range impacts. In other words, seemingly small decisions about which parameters to include or exclude in the models can have a large impact on the outcomes presented – and on the policy debate.

Rather than present a range of outcomes derived from such models, we construct a baseline look at what is at stake. Our goal is to disaggregate the potential regional impacts and to discuss them in two interrelated ways – one that uses the theoretical and empirical underpinnings to gauge potential impacts based on the level of urbanization and another that connects the IPCC estimates to regional socioeconomic data to describe the range and location of potential impacts.

53.2 Expected Impacts Based on Level of Urbanization

53.2.1 Density-Dependant Impacts

Regional differences in population densities bring with them differences in the diversity and extent of economic activities, differences in the need for infrastructure systems and services, and differences in stresses on the local environment. The latter range from the need to convert land to make room for people and their economic activities; to changes in water availability, air quality, and species

diversity; and to changes in the local and regional climate. Global economic and environmental variables have become essential drivers behind these local and regional changes, often exacerbating already existing stressors on social, economic, and environmental conditions.

With most of the world's largest cities located close to oceans, lakes, or rivers, flooding has been and will continue to be a major concern to urban populations. While the average rate of sea level rise from 1961 to 2003 was approximately 1.8 mm per year, that rate has accelerated during the last decade of that period to 3.1 mm (Intergovernmental Panel on Climate Change 2007). As a result, low-lying areas become more readily inundated during high-tide events, storm surges are magnified, coastal ecosystems and their abilities to protect inland areas are rapidly lost, and aquifers and agricultural soils near coasts lose productivity because of salt water intrusion. Recent estimates suggest that by the year 2080 sea level rise and its associated impacts will affect five times more people than it did in 1990 (Nicholls et al. 1999) because both climate change and coastal population sizes will accelerate. Some settlements, particularly many small island communities in the South Pacific, are likely to be completely submerged.

Increases in the frequency of heavy **precipitation events** – most notably rainstorms but also snowfall – have been observed globally throughout the twentieth century, and these trends are expected to continue throughout the twenty-first century. As a result, in some cases, the lakes and rivers on which cities are located will flood; local storm water and flood control systems will be overwhelmed; residential, commercial, and public infrastructures will be inundated and in some cases destroyed (United Nations Human Settlements Programme 2011); ecosystems will be impacted by increased loading of wastes – from untreated sewage to debris to runoff of fertilizers, pesticides, and other potentially harmful substances – and thus experience a loss of their water absorption capacities; water quality will be impaired; land will erode, leaving uphill populations with a need to cope with destruction of their living space and downhill populations with the challenges of dealing with the influx of materials – from soils to debris and other wastes; human physical and mental health will be compromised and lives will be lost; and economic productivity will be undermined, and with it the ability will be reduced to prepare for future flooding impacts because of the need to divert funds for emergency measures and rebuilding efforts and because the loss of economic activity may bring with it a loss of regional, national, and international competitiveness. In some instances, heavy precipitation and rising sea levels, particularly during tropical cyclones, will combine to affect millions of urban dwellers.

Extreme heat events are also predicted to become more frequent and intense. Impacts of heat on urban populations will vary considerably, depending on their acclimatization (Ruth et al. 2006) and their ability to invest in cooling – requiring access to air and space conditioning but also changes in building materials and designs, as well as development of green spaces in cities. Some of these investments, especially where they will require increased energy demand, are bound to contribute to urban **heat island effects** (Akbari 2005), which in turn may set off a spiral of higher energy consumption, further increases in urban temperatures,

changes in regional precipitation patterns, and declines in urban air quality, leading to further cooling needs. Particularly when heat waves coincide with droughts, exacerbation of urban heat island effects and stresses on water and energy supply and distribution ensue. And because not all sectors and households in the urban environment have equal need for or access to cooling, water, energy, disaster relief, and health services, climate impacts in urban areas will likely be not uniform and often exacerbate already existing economic and social inequalities.

Urban areas, of course, are intricately linked to their hinterlands through exchange of water, energy, agricultural, and manufactured goods, services, and people – commuting or migrating, sending and receiving money, or setting trends and expectations. As a consequence, impacts of climate change on urban areas are likely to ripple through to affect larger regions and potentially the global flow of people, goods, and services, and vice versa, and impacts on rural areas will make their mark on the economic, social, and **environmental performance** of cities. About half of the world's population lives within 200 km of a coastline (Small and Cohen 2004).

One notable example of urban–rural interconnections concerns the provision and use of ecosystem services – from flood control to provision of building materials to supply of food and beyond. Such services are essential to revenue generation and quality of life in cities. However, already approximately 60 % of ecosystem services evaluated in the Millennium Ecosystem Assessment are considered degraded or used unsustainably (Millennium Ecosystem Assessment 2008). Increasing urbanization and climate change are likely to continue undermining the provision of ecosystem services with far-reaching consequences for local, regional, and global sustainability.

53.2.2 Agriculture and Forests Impact

Opportunities to study and understand potential climate impacts on agriculture and forests are provided both through natural experiments – such as the El Niño/ Southern oscillation and North Atlantic oscillation phenomena – and deliberate manipulations through the so-called **FACE experiments** (free-air CO₂ enrichment experiments). The latter allow for inferences from temperature variations, while the former test impacts of enriched carbon environment on growth of plants and forests. One of the main findings is the difference in responsiveness between two types of respiration systems found in plants. Most herbaceous plants have a C3 metabolic pathway, including wheat, barley, oats, rice, and soybeans. Higher carbon dioxide levels enhance the growth of these plants. Corn, sugar cane, sorghum, millet, and tropical grasses are C4 plants. For these, higher temperatures (to a certain threshold) are beneficial, but they do not appear to respond to increased atmospheric CO₂. Another physical response mechanism involves a long-ago developed adaptation technique in plants. In times of decreased water availability, the stomata on leaves close, reducing **evapotranspiration** and, therefore, the plant's water demand. In fact, although some studies indicate that the overall irrigation requirements will greatly increase in the USA in the upcoming decades, stomatal closures reduce the impact by around 35 %. FACE experiments with forests indicated that young trees

respond very well to increased CO₂, although mature forests have a much lower (and in some cases, negligible) response (Backlund 2009). Clearly the distribution of plants with differing metabolic systems within and across continents is one component in identifying regional differences in risks.

An aspect of plant-level processes that adds variability to impacts of climate change on agricultural production comes from the different responses that plants show to changing ecosystem attributes, such as shifts in conditions favorable to insect populations, weeds, and disease agents. For instance, although soybean yields are projected to respond positively to higher temperatures, one experiment showed that damage from a harmful insect increased by 57 % when fed on soybeans grown under higher CO₂ concentrations. The main unknown effect is the result of the competition between C3 and C4 plants. Many weeds are C3 plants, so their ranges will likely spread in the upcoming decades. On the other hand, the most popular herbicide, glyphosate, has been shown to lose efficacy for plants grown with more CO₂. Another costly (but unaccounted effect) is the projected higher demand for nutrient inputs.

Although water will be the **limiting factor** in some regions in the following two to three decades, productivity of forests and many crops (wheat, corn, soybean, less so for cotton) is expected to increase for other regions. In fact, timber productivity may increase from 20 % to 60 % in the next three decades. Beyond a certain temperature threshold, however, many crops will be unable to survive. If greenhouse gas emissions continue unabated, we will likely reach this point around 2050–2060. The concept of these thresholds is a bit misleading, however, since most productive and marketable yields are produced under much lower-temperature conditions. For instance, even though the official temperature threshold for corn is around 35 °C, optimal yields are achieved between 18 °C and 22 °C (Backlund 2009).

53.2.3 Natural Landscapes

Grasslands, inlands, and coastal wetlands, shrub ecosystems, and other hotspots for ecological biodiversity will likely experience declines in area and probable changes in their functionality. The more fertile ecosystems with quick reproduction and decomposition rates may initially benefit from warmer temperatures and increased CO₂ concentrations. Yet more severe storm events will likely counteract these benefits as greater **runoff** rates contribute more nutrients to the local waterways stressing aquatic ecosystems and water resources. A comprehensive review of 866 studies on movements of ecological patterns found that several worrisome changes can be attributed to warming. Particularly at risk are range-restricted species and mountaintop species that will see drastic contracts of their ranges, tropical coral reefs and amphibians are also affected, and many disruptions to coevolutionary forces between predators and their prey as well insects and plants have been observed. Impacts on migratory and songbirds, butterflies and dragonflies, flowers like lilac and honeysuckle, and aquatic and tropical species are already apparent (Parmesan 2006).

In most cases, **ecosystem resilience** is being tested not just from the changes in climatic factors but also by other challenges associated with expansion of human activity, such as fragmentation of suitable habitats when forests and grasslands are converted to agricultural or urban uses, chronic overuse of fertilizers, discharge of pollutants, interference with water ways, and deliberate or incidental introduction of alien species.

Beyond plant and animal **biological responses** to changes in levels of greenhouse gases or temperatures or surrounding moisture in the atmosphere, land areas with agricultural production and forest lands are also affected by damages from more frequent severe weather like massive flooding and extreme heat and drought. Recent examples of incidents that illustrate severe weather impacts include a series of heat waves in Europe in 2003, 2007, and 2010, immense landslides in South America and Asia in 2010 and 2011, the sweeping fires raging across vast swaths of Russia and Australia in 2010, and the powerful and devastating floods in Pakistan the same year. The impacts to urban infrastructure, agricultural fields, rural landscapes, and people's livelihoods are clearly immense.

53.3 Regional Differences in Risk and Mitigation Capacity

Some impacts from climate change will likely be similar across economic sectors in many countries around the globe. One useful way to understand such impacts is by differentiating them according to regional levels of urbanization. The severity of climatic effects further depends on a region's **physical geography** as well as regional and local economic and sociopolitical arrangements. For example, diversity in economic sectors, levels of investment in assets at risk from climate change impacts, and dependence on **vulnerable infrastructure** define economic threat. The built-in capacity of institutions to plan, respond, adjust, remain flexible, innovate, and cooperate across government offices and across international borders is inherently local and underlies the duration and magnitude of the impacts.

The stressors on systems are similar across the globe: increases in temperature, changing precipitation patterns, rising sea levels, and more frequent intense storm events. The resultant effects determine shifts in species diversity and distribution, functional changes in natural and managed ecosystems, changes in water availability and pathogen transport, and disruptions to human-made infrastructure. The affected sectors are likewise similar. Agriculture, forestry, tourism, hunting and fishing, coastal real estate, insurance, as well as physical property like buildings, bridges, roads, railroads, and airports may see damages or disruptions as climate change intensifies. Estimating impacts requires measuring many types of activities at their location – settlement and infrastructure sensitivity, food security and agriculture, ecosystem sensitivity to disturbances, human health sensitivity, and water resource sensitivity. The extent of these losses depends, first of all, on exposure, that is, the distribution of assets prone to risk. On that front, stark regional differences emerge.

The Intergovernmental Panel on Climate Change (IPCC) provides the most comprehensive review of regional impacts (Intergovernmental Panel on

Climate Change 2007). For the purposes of the 2007 review, the IPCC uses its own regional designations to describe patterns of impacts estimated through several different global climate models. We match those designations with data from the World Bank and the Food and Agricultural Organization of the United Nations to sketch out the factors that form the baseline for regional differences in terms of impacts and in terms of capacity to respond. The snapshot we provide is from the decade after the international community started negotiating a united policy framework to deal with contributing factors to climate change in 1990s. We use the latest data available for the 2000 to 2010 period (details are described in the Appendix at the end of this chapter) and connect it to the climatic impacts the IPCC projects for the same regions. This aggregation serves to connect the socioeconomic layer to climatic projections. Yet it conceals many of the country-specific and within-country differences, as well as the ever-evolving status of socioeconomic indicators. Still, the aggregation allows for a regionally comparable look at the **baseline conditions** that already underlie the vastly different response capacities – both to mitigate and to adapt to the projected changes. Understanding the starting point onto which such future changes are projected is essential to understanding the entire scope of potential impacts. [All monetary figures used are in US dollars].

The map below shows the regional designations used here.



53.3.1 North America

The continent of North America as a whole is projected to see more weather-related storms of all types and increases in associated damages. But in contrast to the western, southeastern, and northeastern regions, which already have incurred costs related to the changing climate, the northern portions of the continent may initially experience benefits as milder winters bring longer growing seasons. A study of potential economic impacts from climate change in the United States based on eight

regions revealed the depth of those differences (Ruth et al. 2007). Looking at the snapshot of the socioeconomic situation in 2000s shows that around a third of the US population resided along all the coasts. But the population was not distributed evenly. A third of the country's private property was located on the northeastern coast, for example, which is also home to four of the largest cities in the United States. The eastern side of the continent is especially prone to impacts from rising sea level because a natural process of **subsidence** is already pulling the continental board downward. The northeast portion of the country has already seen an increase in severe weather, with the largest increase in very severe events, complemented by a warming of 2.2 °C. An insurance company's analysis found around US\$4 trillion in assets vulnerable to hurricanes in that area alone. What are expected to become more common, category 4 hurricanes touching down in heavily populated metropolitan regions could cost upward of \$50 billion (Ruth et al. 2007).

Major American cities have points below sea level, such as New Orleans, Miami, Jacksonville, Houston, Boston, New York, Washington DC, and Seattle. Five of the ten cities most exposed to a 1 in a **100-year flood event** are here (Nicholls et al. 2008). The estimate of a flood occurring of that magnitude is based on historical data, and projections indicate that such events will become more and more frequent. All together, around US\$19 trillion of insured property are potentially on the path of North Atlantic hurricanes. Since sea-surface temperature plays a major role in hurricane formation, scientists are exploring the possibility that climate change may indeed intensify storms. While it is still a matter of some debate, a recent study found a dramatic shift in the average annual number of tropical storms and hurricanes between 1995 and 2005. The previously steady rate of 9.4 storms jumped by over 50 % to reach an average of 14.8 storms per year (Pearce 2005; Hecht 2007). States on the southern border of the continent – Texas, Alabama, Georgia, Florida, and North Carolina – have each seen over 20 natural disasters causing damages over \$1 billion in the 25-year period between 1980 and 2005 (Lott and Ross 2006). In the United States in 2000s, hurricanes have caused an average damages of \$5 billion per year (National Oceanic and Atmospheric Administration 2010). Nearly 90 % of the population of the continent lives in urban areas, implying that impacts that affect urban infrastructure may be more immediately relevant. The continent enjoys a vast transportation network with around 730 million kilometers of railroad and nearly 8 million kilometers of roads, most of them in the United States.

The southeastern, southern, and western regions of the continent will see severe challenges related to water availability. In addition to a complex political system that guides current water distribution in much of the western portions of the United States, climate change will bring a much drier climate there. Not only water resources for human consumption will be stressed but also water needed to sustain the natural ecosystems and fauna.

The center of the continent relies much more on agriculture. Many millions hectares of agricultural crops that are used not only domestically but are exported throughout the world are grown in the United States. But this means that changes in weather patterns, especially more frequent extreme weather events, cause much

damage. Flooding in the Midwest – which has become more frequent – causes billions of dollars at a time. For example, floods in the summer of 2008 caused \$15 billion damages in the region. Crop damages totaled over \$2 billion that year. On average, floods in the USA caused \$5 billion annually in 2000s. Over the last 10 years, 15 % of flood damages were to crops, one of the most important agricultural commodities. The United States Global Climate Change Impacts team projects continued increases in precipitation and flooding in the center of the continent. Agriculture will see changes across the country. Overall, heat stress will likely alter the relative composition of pests to plants to nutrients used. Increased insect outbreaks are possible in the northwest reaches of the continent to the southwest (Karl et al. 2009). Temporarily, climate change may extend growing season and be beneficial toward the industry, such as in the Pacific Northwest. But other concerns may undermine this trend. For instance, many invasive species benefit from warmer climates, and they cost \$120 billion a year to control in 2000s (Pimentel et al. 2005).

Although not many people work in **agriculture**, the sector is very important economically and geographically. Across the entire continent, agriculture added over \$50 billion to the economy annually in 2000s. In the United States, agricultural uses take up around half of the country's land area. Recently, studies show adverse impacts from climate change on specific agricultural industries across the nation – for example, the dairy cows suffering lower productivity as temperatures rise or grape quality diminishing as springtime advances. Milk production and wineries are small but growing profitable agricultural activities, especially in the United States. As research continues and focuses on more specific industries and locations, more impacts are revealed.

53.3.2 Europe

The European continent will see varying changes. For example, every scenario run through the models used for the latest IPCC report indicate that the northern areas of Europe will see greater warming during the winter months, while southern regions will need to prepare for hotter summers. Maximum temperatures experienced over an average year are projected to rise more steeply in central and southern European countries. The socioeconomic context in the 2000s provides the baseline conditions on which the impending changes will occur. A quarter of Europe's 600 million people resided in one of the southern European countries and the Mediterranean region – Italy, Portugal, Slovenia, Serbia, Albania, Greece, or others. Another quarter lived in eastern Europe. On average, projections show increases in precipitation in the north, but decreases in the south – albeit intensity of the events will continue to strengthen across the entire continent. Incidence of **heat waves** and duration of droughts will increase in central and southern European countries. Countries in western Europe and alpine regions will also see more dry periods and hot days. Melting of **permafrost**, less snowfall, and loss of glaciers in mountainous regions are expected, as well as increased flooding along the coastlines. In general, changes in precipitation will affect water resources, with consequences for

both the non-managed and managed landscapes. Since less wintertime precipitation will end up as snowpack, flows in major European rivers during winters will increase, while decreasing in summer months. Eastern and southern Europe may see especially dramatic declines. Because of regional climatic trends, sea level rise along the European coasts may be 50 % more than the expected global average. In 2009, the European Commission's Directorate General for Maritime Affairs conducted a survey of 22 coastal European countries, finding that around US \$700 billion to 1.4 trillion of assets were located within half a kilometer of a coast. Over a third of the GDP of these countries is created within 50 km of the coastline (Directorate-General for Maritime Affairs and Fisheries 2009).

Infrastructure impacts reverberate through many sectors. For instance, transportation reliability may suffer. In 2000s, railroad systems in eastern and Western European countries transported 95 billion and nearly 220 billion passenger kilometers on an annual basis, respectively. The total network of roads was over 2 million kilometers in western Europe and around 1.5 million kilometers in the other regions. But again, distribution matters. On average per country, road density in Western European countries was nearly eight times that of Eastern European countries and four times of southern European countries. The exposure portfolios are starkly different. To every 1.16 cars that a person in western Europe owns (or 1.5 in North America), an Eastern European had a third of a vehicle. This means practical differences in the level of development of infrastructure and, therefore, its exposure to impacts. It also means that personal mobility of families may be compromised in an emergency.

Because many of the trends in climate changes are ongoing, researchers have observed impacts on those systems already. The capacity to deal with the impacts relates to how exposed countries' portfolios are to particular risks. For instance, managed landscapes like **croplands** and **fisheries** will be at risk from a combination of factors, but the resultant damage may be less severe as affected parties adapt to changing conditions. The associated socioeconomic impact of that response comes down to several factors. To a large degree the impact will depend on the amount of water available to cope with increasingly drier conditions. In southern European countries, about 10 % of agricultural lands were irrigated – in contrast to less than 5 % in northern Europe. On average, southern and Eastern European countries were economically more dependent on agriculture than the other regions. Employment was more concentrated in agriculture too – with about a sixth of the population working in agriculture in southern and eastern Europe in 2000s. Changes in the climate threaten agricultural stability, potentially threatening economic livelihood of many people. On the other hand, on average, agricultural employment stood at 3 % in Western European countries and at less than 5 % in northern Europe. The southern and eastern regions appear to be more vulnerable. For example, although mild warming will have little effect on agriculture, increases of over 5 °C can lead to 10 % reductions in crops overall, but around 25 % reductions in southern Europe (Agrawala 2007).

Water availability and flow affect capacity to produce **hydroelectrical** power. In 2000s, countries in northern Europe derived about a third of their total electricity production from such sources. Over a quarter of electricity production in southern European countries came from hydropower. Eastern European countries annually

utilized over half of their internal freshwater resources, in contrast to Western European countries using a third and southern European countries about a quarter on average. This is in comparison to less than 10 % used by countries in North America, signaling some strain on the resource already across the entire continent.

53.3.3 Asia

The Asian region spans across Russia to the Middle East through Southern Asia to China to the Pacific coast and its many islands in the IPCC report. In 2000s, around two-thirds of the world's population lived here. While the number of rainfalls in the region has declined, the severity of storms has gone up. Severe storm events are now more frequent and more intense, resulting in increased floods and landslides. This trend is expected to continue as average precipitation is expected to increase across this sweeping region, especially across boreal forests in Asia. But parts of the continent that are currently arid or **semiarid** – most notably regions in Pakistan, India, and Indonesia – will continue to experience decreases in rainfall and see an increase in the number of **droughts**. In a similar manner, the number of tropical cyclones in the Pacific has dropped, but the intensity and the resultant damages of cyclones that form have gone up. The duration of **heat waves** has prolonged and will likely continue to prolong across Asia. Annual warming of 3 °C by the 2050s and of 5 °C by the 2080s is projected on average. Highest warming rates have been observed in North Asia, including Russia. Once again, risk to species and entire ecosystems is intensifying, altering functional relationships and pushing their physiological boundaries.

In 2009, the region was home to 4.2 billion people, with about 40 % living in southern Asian countries like India, Bangladesh, Pakistan, Iran, and Afghanistan. Nearly as many people lived in eastern Asia – in China, Japan, South Korea, or North Korea. Another fifth of the population resides in southeastern Asia, in Indonesia, Thailand, Vietnam, Fiji, or the Philippines. Many of these countries are **small island nations**. These are especially vulnerable to sea level rise. Six out of ten coastal cities with the highest populations vulnerable to a very major flooding event are spread across Asia (Nicholls et al. 2008). One study on three megacities in Asia – Manila, Ho Chi Minh City, and Bangkok – found that by 2050 damages associated with **floods** will be more and more substantial – 2 % to 6 % of the region's GDP. Much of the damage was attributable to the expected land subsidence (The World Bank 2010).

Intensified **water stress** can reduce crop yields of essential diet staples like rice, corn, and wheat. Yields for rice crops decline 10 % for every degree Celsius increase. Area of land suitable to agriculture is projected to decline in east Asia, which now has 16 % of its land classified as arable. Russia and countries in central Asia will likely see expanded agricultural production, yet it is uncertain how different crop portfolios will react to projected changes. For instance, China grew more than 30 million hectares of corn and nearly as many of rice. Russia and Pakistan had 25 million hectares of land under cultivation for wheat production each. In 2000s, agricultural production took up large swaths of the land.

For instance, around 44 % of land in northern Asia agriculture and 60 % of land in central Asia were under agricultural cultivation. Western Asian countries like Saudi Arabia, Iraq, Turkey, Israel, United Arab Emirates, Georgia, Armenia, Lebanon, and others had over a third of their land in agricultural production, on average. The agricultural sector was much more important to the region's GDP and the population's employment than in Europe or in North America. For example, in 2000s, the agricultural sector contributed nearly a fifth of value-added activities to the GDP in central and southern Asia, where around 40 % of people were employed in agriculture. Clearly, many people will be exposed to climate change impacts to agriculture.

Plus, the same impacts threaten natural ecosystems. Asia is home to some of the most biologically rich spots, like the ones in China, Japan, Russia, India, Indonesia, and Papua New Guinea. To a large extent, such biodiversity continued to prosper because considerable portions of the continent were left undeveloped. For example, the infrastructure of the region in the 2000s decade was less extensive than in Europe or in North America. Around 40 % of the roads were paved in southeastern and northern Asia, each with about a million kilometers of total roads. Southern Asia had about 5 million kilometers of roads, many of them unpaved. The Asian continent – both with its megacities and sizeable rural populations scattered across the landscape – may be more vulnerable to certain climate change impacts because of its socioeconomic portfolio more so than the regions discussed thus far.

53.3.4 Latin America

The Latin American region, as delineated by the IPCC, stretches from the Caribbean and Central America in the north to the very southern tips of Chile and Argentina containing around a tenth of the world's population (about the same percent as Europe) in 2009. The IPCC reports that most of the central belt of the South American continent has seen an increase in precipitation, although southern Chile and regions up along the western coast – southwest Argentina and southern Peru – have observed lower levels of rainfall. **The Amazon** has seen a 10 % increase in flood frequency, and rivers in the center of the continent have had a 50 % increase in their **streamflows**. Little data exist about the middle of the continent, making it difficult to discern any trends. But much is known about the glaciers, which are receding at an accelerating pace as temperature and humidity conditions change and precipitation cannot compensate for the rate of melting. The **glaciers** spanning the continent are projected to disappear by the mid-2020s. Not only is this a loss of an important ecological constituent, the disappearance of the beautiful skiing slopes threatens an important industry and limits recreational opportunities.

Models predict warming for Latin America on average with increases in temperatures from 1 °C to 7.5 °C by the end of the century. The occurrence of significant storms and periods without precipitation will likely increase. Frequency and intensity of hurricanes around the Caribbean islands will also likely increase – the 2001 and

2005 seasons were two of the worst on record. The United Nation Environment Programme estimates that over 11 million individuals in Latin America were affected by natural disasters in 2001 and in 2005. Around 530 million people lived within 100 km of the coast in 2005 (United Nations Environmental Programme 2008).

In the central portion of South America, agricultural production has seen a benefit from greater precipitation – soybean yields increased up to 38 % and corn to 18 %. But natural land cover is retreating as tropic **deforestation** continues, fueled by the booming prices of agricultural crops like soybeans and corn (although notably by 2012, the rate of deforestation fell). In 2000s, agriculture took up nearly a third of land area in the Caribbean and South American countries. It also contributed to roughly 10 % of the GDP of those regions. Large-scale crop production was not widespread in the Caribbean, where only Cuba, Haiti, and the Dominican Republic had sizable plots of rice and corn. Expansive countries like Brazil and Argentina had around 20 million hectares of soybeans each. Climatic changes affect agricultural production and agricultural prices. Future planting decisions may need to take into account the agricultural commodity's resistance to a host of climatic variables. On the other hand, as mentioned, many areas of the continent have seen positive effects from the climatic changes. Plus, much of the region has plentiful water resources, sparing it the impacts associated with water availability.

In contrast to some other regions, in 2000s, South American and Central American countries used their internal water resources sparingly, withdrawing 2.4 % and 4.4 % of their freshwater annually. Still, other areas faced a different situation. Caribbean nations withdraw more than 20 % on average, making them more vulnerable to upcoming changes in water availability. Plus, hydroelectric sources constitute the main source of electricity for many Latin American nations.

In 2000s, a large portion of the populations live in rural areas – 40 % in Caribbean and Central American countries and 25 % in South American countries. The difference in infrastructure development has the same implications as discussed before. Less concentration of people makes the region less vulnerable to some of the impacts. Yet there are still incredibly dense areas like in the Caribbean nations where 60 % of the population lived in the largest city. That region is also the one more likely to suffer from increased frequency in extreme weather events. There, much infrastructure is at stake since road density in the Caribbean nations is than 12 times the road density as the open South American continent. Nonetheless, generally speaking and as a whole, the continent seems to be fairly resistant to the most serious impacts, at least in the short- to medium-term timeframe.

53.3.5 Africa

The African continent is home to one-seventh of the world's population. The IPCC warns that it will likely be hit the hardest by a combination of changing climatic factors and existing development challenges. The continent is projected to see a 3–4 °C increase in mean annual temperature, but the northern and southern

parts will likely see disproportionate warming with up to 9 °C increase in the north during the summer and up to 7 °C increase in the south during the spring (September to November) by the end of this century. Northern African countries are some of the driest in the world – they received less than 200 mm of rainfall per year in average precipitation, in contrast to the arid Middle Eastern countries that receive 330 mm of rainfall per year on average. Projections for changes in precipitation in Africa are less clear, especially across the **Sahel desert**. Generally, however, precipitation will likely decrease in the northern African countries up to 20 %. It will follow a similar pattern in the south but will increase in the east. It is uncertain which precipitation trends will emerge across the expansive Sahel region. Still, the number of extreme dry and wet years is projected to rise – an expectation consistent across the globe.

Although in 2000s a tremendous shift from rural to urban lifestyles was and has been ongoing across the continent, a large proportion of the population lived in rural areas. It was the largest in eastern African nations with 68 % and the lowest in northern African nations with 41 %. Agriculture was a significant portion of the region's GDP – reaching 32 % in western African nations and 25 % in Eastern Africa. Employment in agriculture is also high with nearly 50 % of the population working in the sector in western and central African regions. Whereas about half of agricultural cropland in the North American nations is dedicated to just three crops – corn, soybeans, and wheat – much of Africa's agriculture is done on a smaller scale translating into more diversified cropping patterns. Yet as climate changes progressed, agricultural production will change as well. One study indicated that wheat may disappear from the continent altogether. Some positive effects on agricultural crop production are possible in the eastern areas, but northern, central, western, and southern Africa showed negative trends.

Libya, Sudan, Egypt, and other northern African countries are facing severe water shortages. The three countries withdraw much more freshwater resources than is available already. For the 250 million people in the northern and southern African regions whose water resources are already stretched and for whom climate change will likely mean a decrease in precipitation, the situation will worsen. Western and eastern African nations, however, may see some relief. In 2000s, the regions annually withdrew 12 % and 34 % of their **internal freshwater resources**, respectively. This could mean that enough water may be available to continue the central and eastern African regions' reliance on hydroelectric sources to provide electricity to its populations with nearly 80 % and 63 % of the total production generated by these sources. In contrast, northern and southern African nations got 8 % and 23 % of the electricity from these sources.

The road network was not as extensive in Africa as it is elsewhere. This indicates that movement may be difficult in an emergency. In comparison to road density in North America and Australia of about 40 km of road per 100 km of land, road density ranges from around 8 km of road per 100 km of land in central Africa to 28 km in the eastern countries. Many of those kilometers are not paved, however. About 15 % of the roads were paved in central Africa, 20 % in western Africa, and 66 % in northern Africa. Vehicle use there was the lowest out of all the world

regions. This cursory overview does little justice to the extent of the social, developmental, and political challenges many regions on the continent face. Impacts from climate change, unfortunately, appear to be another challenge for many areas. The damages from the impacts may be magnified because of inadequate capacity – structural and political – to reform and to adapt.

53.3.6 Australia and New Zealand

Australia, New Zealand, Tasmania, Fiji, and Samoa are projected to experience significant warming by the end of this century. Temperatures in the central portions may increase up to 8 °C with smaller increases closer to the coast – up to 5.4 °C within 400 km of the coast. Near the coast precipitation is expected to drop by up to 80 %. Southern and subtropical Australia will also see decreases, but northern and central territories may see some increases. Just like on the North American continent, southwestern regions of Australia will see many more droughts – up by 80 % by 2070 as simulated by one study. New Zealand will likely also experience more frequent severe droughts. Precipitation will increase in the west and decrease in the east.

In 2000s, out of the population of 27 million people (or 4 % of the world's population) in the region, 85 % of Australians live within 50 km of the coastline, nearly 100 % of Tasmanians and all residents of the small islands. Everyone in New Zealand lives within 100 km of the coast. This region of the world is especially vulnerable to the impacts from climate change related to encroachment of seawater and more frequent storms and hurricanes.

Australia, New Zealand, and **small island nations** are projected to experience particularly harsh impacts from climate change – some of these like changes to natural ecosystems and stress on available water are already underway. Economic damages from consequences of more frequent natural disasters, such as floods, droughts, storms, or landslides, are rising. The 2002 drought cost Australia \$7.6 billion. New Zealand alone suffered losses to its agricultural sector of \$800 million in the multiyear droughts in the late 1990s. Floods cause an annual \$85 million in damages there.

The populations of these nations were concentrated fairly tightly with nearly half of them living in the largest city. Australia and New Zealand's rural population was about 10 % of the population, while Fiji's was closer to 50 % in 2009. The road density in Australia was close to that in South America or central Asian countries like Kazakhstan or Uzbekistan. The stretch of the road network was six times smaller in Australia and New Zealand than in the United States. This means that less infrastructure is on the path of destruction, yet mobility may be more limited. The fairly sizable rural populations also mean that agriculture is an important source of income for many – and more comparable to the numbers for Europe and North America.

On average, 8 % of the continent's GDP was attributable to the agricultural sector, which employed around 5 % of the population in 2000s. Agricultural productivity may be compromised as a result of temperature increases and changes

in precipitation. In 2000s, Australia received about 530 millimeters of precipitation each year – less than three times what New Zealand got. This is comparable to the average on the southern tip of Africa. Future projections for lower precipitation across much of the continent puts in question continued expansion of agriculture (which grew nearly 7 % in Australia, but contracted 12 % in Fiji according to the latest available numbers). Lower water supplies may also alter how much electricity is derived from hydroelectric sources. While Australia derives just over 5 % from hydro, New Zealand draws more than 55 % from the source. Overall, the continent has strong baseline conditions to deal with many projected impacts in the short-term future, but its capacity may be strained as impacts intensify.

53.4 Interconnectivity and Reach of Impacts

The economies of the regions described above are highly interconnected globally. This **interconnectivity** can propagate environmental impacts in one region to affect many others – one poignant example of such was the quick shrinkage of sales of Japanese automobiles across the globe when production halted following a disastrous tsunami in the spring of 2011. The magnitude of such **ripple effects** depends on the extent to which a country's economy is tied to global markets. International tourism is a particular case in point. This service industry is the economic backbone of many beach destinations, skiing and winter sport spots, and wildlife-touring locations. In 2000s, tourism attracted over 70 million people from around the world to the North American continent annually, generating over \$100 billion in receipts. And travel services account for over a quarter of the countries' average commercial services exports. Uncomfortably high temperatures, lack of snow, and degradations of ecological landscapes can diminish the sector's contribution to national economies.

Several European countries share the fears of declining tourism because of climate change, especially to the well-off snow sports industry hugging the Alps, whose annual profit is on the order of \$70 billion. Only a third of the current number of resorts will remain if warming reaches 4 °C, and those able to remain open may need to supplement snow through artificial means, raising their operating costs and reducing their competitiveness. In 2000s, southern economies were also dependent on international tourism, which accounted on average for 20 % of the total export bill. All together Western European countries hosted 150 million tourists annually, while southern European countries welcomed 140 million. Interconnectivity is strong not only through travel but also through **trade**. This is especially true for the Western European countries where two-thirds of their GDP came from exports. A third of southern European countries' GDP related to exports.

Asian countries' economies are also closely tied to trade relationships with other nations. In 2000s, exports were growing at very high rates in many Middle Eastern countries – like Qatar, Oman, Turkey, Saudi Arabia, Afghanistan – and countries previously tied to the Soviet Union: Turkmenistan, Tajikistan, Georgia, Armenia, and Azerbaijan. For eastern Asia, exports accounted on average for three-quarters

of the countries' GDP. International tourism was booming. All of Asia welcomed around 330 million tourists in 2000s – or nearly 30 % of the total world traffic – and brought in over \$265 billion. Eastern Asian countries, especially Hong Kong, China, and Korea, are particularly active in the global stock market, trading annually in stocks three times the value of their GDP, on average.

Latin American countries also leaned heavily on exports to support their economies, realizing also that international tourism was a profitable business. It brought in around \$60 billion for the Latin American region with over 6 million annual visits in 2000s. Australia and New Zealand economies rely less on trade but have a lucrative international tourism industry. Although the region attracted less than 1 % of world's international tourists with around 8 million visitors, it drew 3.2 % of the tourist receipts.

The situation in Africa in terms of exports is much different than in the rest of the world. Although trade constitutes a large portion of most African nations' GDP, **export and import volumes** for the continent were relatively low during the same time frame. For instance, the value of the Africa's imports and exports was a tenth of Asia's. African exports were growing in the eastern and western regions but stagnated in central and northern Africa, even though those areas were net exporters of energy. Africa drew in 60 million tourists a year – about the same number as Latin America or North America. North American countries, however, received twice as much revenue from their guests than their counterparts in Africa.

The variable – and ever-changing – interconnectivity of the world is a difficult dimension to account for when estimating potential impacts of climate change. Each region has a unique suite of economic, social, and historical relations with other nations and experiences a different mix of climate impacts. The baseline conditions in 2000s described above give some perspective on where the different regions were in their economic vulnerabilities. The capacity to deal with the consequences is another dimension that complicates the issue of climate change.

53.5 Global Social Justice

While every continent will see impacts from climate change, some regions are much better prepared to deal with the consequences. Countries on the North American continent for instance have a very high average per capita GDP of nearly \$50,000, affording them the financial means to address some of the consequences. Virtually the entire population enjoys modern **sanitation** facilities and has access to clean water and sufficient food. In 2000s, population growth was less than 1 % per year across North American countries. Resources and time are on the regions' side. Still, challenges remain. Aging infrastructure across the United States is a growing pain. In 2009, the American Society of Civil Engineers split the country's infrastructure resources into 15 categories, giving only four categories the grade of C or C+, while the rest received D or D-. Maintaining **drinking water** requires an additional \$11 billion in annual funds. At least \$100 billion is needed to update the

nation's levee system. The government spent less than 40 % of the \$190 billion needed to preserve the many kilometers of road.

Many European countries likewise enjoy solid baseline conditions to deal with climate change impacts. But sharper differences emerge region to region. GDP per capita ranges from a low of \$9,000 in current US dollars on average in Eastern European countries to a high of \$78,000 in Western European countries. Economies are diversified and the populations are well educated for the most part. This gives some economic stability and capacity. Challenges abound, however. Southern European countries have an average unemployment of over 16 %, for instance, and their workers attained lower educational levels than their counterparts in other parts of Europe. Families with low incomes have fewer resources to deal with the effects of storms or increased temperatures. The cost of cooling equipment and energy can be prohibitive. **Migration** is also high into western and northern European countries, where nearly 3 million people migrate annually. Even more, 3.5 million people, moved to southern European countries, particularly to Spain and Italy. An increase in people stretches institutional resources, perhaps leaving less for adaptation or mitigation purposes. Migration is a hot-button issue in Europe, and one concern there is of an increase in migration as people move to avoid the worst impacted areas.

Countries in the Asian region span a full spectrum of economic and political developmental stages. As a result, government capacity to respond to costly emergencies varies. Health infrastructure, **food security**, and capacity of physical systems to handle growing populations and growing demand for a higher standard of living also vary. In 2000s, GDP per capita ranged from very low ranges (an average of less than \$2,000 current US dollars in southern Asian countries and around \$2,800 in central Asian countries to the average) to medium levels (\$5,100 on average in northern Asian countries and \$6,500 in south-eastern Asian countries) to relatively high incomes of \$18,000 in western Asia and \$26,000 in eastern Asia. Clean sanitation facilities were available to 60 % of the people residing in southern Asian countries, 67 % in southeastern and northern Asia, and 80 % in eastern Asia. Clean water was available to over 83 % of the population in every country, on average. Such present-day challenges undermine the regions' capacity to respond to impacts from climate change. Sanitation and availability of clean water are crucial tools in preventing the spread of **disease vectors**, many of which may benefit from warmer temperatures.

Latin American countries have varying availability of infrastructure. While they still have many unpaved roads, over 90 % of the population has access to good water sources and around 80 % to clean sanitation. The GDP per capita there is comparable to medium-range incomes in Asian countries. Individuals in the Caribbean Basin have around \$10,000 in current US dollars, while those in Central America have less than \$4,500. Individual capacity to respond to disaster depends in part on the availability of resources.

Many of the African nations were impoverished in 2000s. The western African region has the lowest GDP per capita in the world with \$836. Eastern African nations average a GDP per capita of about \$1,500, nearly \$400 less than southern

region in Asia, which has the lowest GDP per capita there. The northern African region has the highest GDP per capita of the continent with around \$4,000 – less than half of what it is in Eastern European nations and nearly 20 times less than what it is in Western European nations. This discrepancy points not only to the different level of economic development but also to the potential individual capacity to deal with a crisis. Adequate health and environmental resources are also necessary to expand individual capacity. About a third of the population lacked access to clean water in central, eastern, and Western African regions. Three-quarters lacked access to clean sanitation facilities in western Africa, and two-thirds lacked access to sanitation facilities in central and western African nations. Northern and southern regions were better off – 80 % of people in the north had access to both clean water and sanitation. Half of the people have access to sanitation facilities, and 86 % have access to clean water in the south. So although actual impacts from climate change will affect northern and southern African regions disproportionately, some capacity exists to adjust.

Australia and New Zealand have capable and stable governmental, economic, and physical structures. The GDP per capita is nearly \$20,000 on average between New Zealand and Australia, but the GDP per capita in Samoa and Fiji are about a tenth of that. Almost the entire population has access to clean water and clean sanitation facilities. There is virtually no **malnutrition** prevalence, whereas rates were at around a quarter of children under 5 or more in all of the regions in Asia (in southern Asia the rate is 41 %, as measured by the height for age ratio) and around the same in Africa – with 45 % of the children malnourished in eastern African countries. Inability to deal with existing problems highlights the level of institutional capacity. More disruptive weather events promise to test it further.

53.6 Conclusion

This chapter presented an overview of the range and magnitude of resources at risk from climate change and highlighted the overlapping climatic, ecological, and socioeconomic dimensions. No particular impacts are certain in their magnitude or timing. Trends, however, are clearly emerging. Still, uncertainty is part of all projections, especially complex ones like those linking global climate change with regional economic and social dynamics. The scientific portion of connecting natural processes to observed and anticipated impacts is on solid footing. The sociopolitical portion is much less certain, not least because we have few ways to communicate impacts beyond economic terms. At a minimum, economic measures provide a convenient common ground for comparing fairly simple concepts across regions, such as damage to physical coastal property, rising insurance costs, foregone prices of ruined agricultural crops, declining receipts in the tourism sector, or repair bills associated with infrastructure damage. More nuanced concepts like the fairness of associated regional distributions of those impacts, the social costs of human suffering related to increased incidence of disease and disaster, the environmental stress inflicted on the natural systems, the widely differing adaptive capacities, or

the fairness questions that divide the developing and developed nations do not easily fit into standard models of climate impacts and responses. They can hardly be reduced to numbers, let alone to dollars. Yet such nuanced and multidimensional assessments will be needed to better understand regional impacts and guide responses.

Appendix

All economic data used in this report came as country-level indicators from the World Bank's World Development Indicators and Global Development Finance database and the Food and Agricultural Organization of the United Nations. We use the latest available statistic from 2000 to 2010 for individual countries for the World Bank data and the latest available from 2000 to 2009 for the FAO data (World Bank 2011; FAO 2009). The countries are assigned regions according to the regions in the IPCC report and subregions using the UN Geoscheme categories because the IPCC report did not provide a country-by-country breakdown of the regions. We made the following modifications. Russia and Mongolia were placed in the Northern Asia region in line with the IPCC report. This is not a region within the UN Geoscheme. Several small island nations were not listed in the UN Geoscheme and were placed to the closest IPCC region. We also grouped Melanesia, Micronesia, and Oceania islands with the Australia and New Zealand region – although because of dearth of socioeconomic data for these nations, many of them were not included in the analysis. We renamed the Middle Africa region as the central African region for more consistency. All data listed is specified as either the average per country or the total data point for countries with available data for the region. All monetary figures are in current US dollars.

References

- Agrawala S (2007) Climate change in the European Alps: adapting winter tourism and natural hazards management. OECD, Paris
- Akbari H (2005) Energy saving potentials and air quality benefits of urban heat island mitigation. Lawrence Berkeley National Laboratory, Berkeley
- National Oceanic and Atmospheric Agency (2010) Economics of heavy rain & flooding data and products (costs). Resource document. <http://www.economics.noaa.gov/?goal=weather&file=events/precip&view=costs> Accessed 7 July 2011
- Backlund P (2009) Effects of climate change on agriculture, land resources, water resources, and biodiversity in the United States. U.S. Climate Change Science Program and the Subcommittee on Global Change Research, Washington, DC
- Intergovernmental Panel on Climate Change (2007) In: Hansen CE, Parry ML, Canziani OF, Palutikof JP, van der Linden PJ (eds) Contribution of working group II to the fourth assessment report of the intergovernmental panel on climate change, 2007. Cambridge University Press, Cambridge, UK/New York
- Directorate-General for Maritime Affairs and Fisheries (2009) The economics of climate change adaptation in EU coastal areas. Summary report

- Food and Agriculture Organization of the United Nations (2009) FAOSTAT, 2009. Resource document. <http://faostat.fao.org/> Accessed 28 September 2011
- Hecht J (2007) Atlantic hurricane frequency doubled last century. *New Scientist*, August 4
- Karl TR, Melillo JM, Peterson TC (2009) Global climate change impacts in the United States: a state of knowledge report from the US global change research program. Cambridge University Press, New York
- Lott N, Ross T (2006) Tracking and evaluating U.S. Billion dollar weather disasters, 1985 to 2005. NOAA's National Climatic Data Center, Asheville
- Millennium Ecosystem Assessment (2008) Ecosystem change and human well-being: research and monitoring priorities based on the millennium ecosystem assessment. International Council for Science, Paris
- Nicholls RJ, Hoozemans FMJ, Marchand M (1999) Increasing flood risk and wetland losses due to global sea-level rise: regional and global analyses. *Glob Environ Chang* 9:S69–S87
- Nicholls RJ, Hanson S, Herweijer C, Patmore N, Hallegatte S, Corfee-Morlot J, Château J, Muir-Wood R (2008) Ranking port cities with high exposure and vulnerability to climate extremes. Organization for Economic Development, Paris
- Ortiz RA, Markandya A (2009) Integrated impact assessment models of climate change with an emphasis on damage functions: a literature review. Basque Centre for Climate Change, Spain
- Parmesan C (2006) Ecological and evolutionary responses to recent climate change. *Annu Rev Ecol Evol Syst* 37:637–669
- Pearce F (2005) Is global warming making hurricanes stronger? *New Scientist*, December 3
- Pimentel D, Zuniga R, Morrison D (2005) Update on the environmental and economic costs associated with alien-invasive species in the United States. *Ecol Econ* 52(3):273–288
- Ruth M, Amato A, Kirshen P (2006) Impacts of changing temperatures on heat-related mortality in urban areas: the issues and a case study from metropolitan Boston. In: Smart growth and climate change: regional development, infrastructure and adaptation. Edward Elgar Publishing, Cheltenham, UK
- Ruth M, Coelho D, Karetnikov D (2007) The US economic impacts of climate change and the costs of inaction. Resource document. <http://www.cier.umd.edu/climateadaptation>
- Small C, Cohen JE (2004) Continental physiography, climate, and the global distribution of human population. *Curr Anthropol* 45(2):269–277
- Stanton EA, Ackerman F, Kartha S (2009) Inside the integrated assessment models: four issues in climate economics. Stockholm Environment Institute, Working Paper WP-US-0801
- The World Bank (2010) Climate risks and adaptation in Asian coastal megacities. World Bank, Washington, DC
- United Nations Environmental Programme (2008) Report on the Latin American and Caribbean initiative for sustainable development (ILAC): five years after it was adopted. In: Sixteenth meeting of the forum of ministers of Latin America and the Caribbean, Santo Domingo
- United Nations Human Settlements Programme (2011) Global report on human settlements 2011: cities and climate change. Earthscan, London\Washington DC
- Weitzman ML (2011) Fat-tailed uncertainty in the economics of catastrophic climate change. *Rev Environ Econ Pol* 5(2):275–292
- World Bank (2011) World development indicators & global development finance. World Bank. Resource document. <http://databank.worldbank.org/ddp/home.do>

Emily Talen

Contents

54.1 Introduction	1071
54.2 Principles	1072
54.3 Sustainable Cities and Regions	1074
54.4 Implementing Sustainability Goals	1077
54.5 Conclusions	1081
References	1081

Abstract

Sustainability has become a key concept in the quest to define a normative framework for urban and regional development. This chapter presents an overview of what is meant by sustainability first from the regional and then from the city level. Both scales have a long history in the planning domain, but the notion of a *sustainable city* is key to both realms and is the main focus of this chapter. While there is widespread agreement on broad parameters and principles about urban and regional sustainability, there are entrenched debates over implementation. On one level, there are debates over implementation methods, especially the degree to which partial success in implementation is better or worse than doing nothing. More fundamental debates about sustainability involve the distinction between process vs. form and the integration of city versus nature.

E. Talen

Arizona State University, Tempe, AZ, USA

e-mail: etalen@asu.edu

54.1 Introduction

On the subject of urban and regional sustainability, the debate is no longer *whether* cities and regions should be less environmentally harmful and more human scaled but what the specific policy and design responses should be – whether government subsidies and funding priorities, market incentives, new kinds of codes, transportation systems, or urban design schemes are achieving what is needed. Our views of the *sustainable city* and region have evolved from “what is it?” to “how do we get there?” in much the same way that many environmentalists decided several years ago that the debate over *global warming* had been settled, despite the continuing pushback from the other side.

Most urban planners would now argue that, in principle and in broad terms, we know what the *sustainable city* and region is supposed to be and what the economic, social, health, and environmental benefits of it could potentially be. The Wikipedia definition adequately sums up the main objective: “A *sustainable city* can feed and power itself with minimal reliance on the surrounding countryside, and creates the smallest possible ecological footprint for its residents.” Sustainable development at the regional level is often more abstractly defined, broadly attempting to balance equity, economic, and ecological concerns. Berke and Conroy define sustainable development as “a dynamic process in which communities anticipate and accommodate the needs of current and future generations in ways that reproduce and balance local social, economic and ecological systems, and link local actions to global concerns” (Berke and Manta-Conroy 2000, p. 23). With these broad principles in mind, the focus, in the Western world and in the USA especially, is squarely on implementation.

This chapter first reviews the generalized principles of sustainable cities and regions, moving from a broader review toward a more specific definition. It begins with the basic principles and then spells out what those principles might mean for the physical form and pattern of cities. It will become clear that as we move from broad principles to specific design strategies, the degree to which planners agree becomes increasingly strained. The second part of the chapter focuses on the key debates within the literature on *sustainable cities*. Although planners largely agree about what a *sustainable city* should be on a certain level, there continue to be entrenched disagreements about the best approach to getting there.

54.2 Principles

We can start with the meta-principle – the notion of “*sustainability*” from the often-quoted *Brundtland Report* is as follows: “Sustainable development is development that meets the needs of the present without compromising the ability of future generations to meet their own needs” (World Commission on Environment and Development the Brundtland Commission 1987, p. 43). Sustainability involves adopting a lifestyle “within the planet’s ecological means” to ensure that development does not compromise the needs of future generations and to ensure that

population growth is “in harmony with the changing productive potential of the ecosystem.” Urban and regional planners have translated this to mean that cities must endure environmentally, economically, and socially, balancing what have come to be known as the three “E’s: environment, economy, and equity (Berke 2002, p. 30; see also Campbell 1996). Basically, this means that planners should help cities develop in ways that last.

Often sustainability is cast as a continuation of environmental planning. While planners seem to agree that *sustainability* requires a holistic view “that includes equal concern for environmental, economic, and social sustainability” (Daniels 2009, p. 185), the environmental perspective dominates. Thus, urban *transport* is to be energy efficient, solar power is to be promoted where possible, and water is to be used efficiently – in short, cities are to be redefined as “eco-technical systems.” Cities are going “green,” and thus Routledge’s comprehensive four volumes entitled “Sustainable Urban Development” are principally devoted to environmental assessment. Out of this larger environmental focus, many subtopics have evolved that are of particular relevance to urban planners, like the relationship between sustainability and technology, sustainability and architecture, and even “sustainable Olympic games.”

Sustainable development can engage a complex array of political views that range from “free-market” *environmentalism* to ecofeminism, animal rights, and bioregionalism. The three-way conflict between *environmentalism*, economic development, and social justice – green cities, growing cities, and just cities, as Campbell (1996) refers to them – is present in all of these approaches, and each manifests a human vs. nature duality to varying degrees (this duality is discussed in more detail at the end of this chapter). Proposals include “greening the market,” liberal *environmentalism* in the tradition of John Rawls, ecosocialist theory, or the biological rooting of culture through “reinhabitation.” In many of these applications, there remains a fundamental, lingering duality that conceptualizes an environmental crisis in human vs. nature terms.

While these concepts have found their way into the rhetoric of metropolitan development reform, there is a significant question about the degree to which rhetoric is being translated into actual practice. Sustainability is a concept endorsed by both economic development proponents and radical ecologists, and as Campbell (1996) points out, “any concept fully endorsed by all parties must surely be bypassing the heart of the conflict.” Cultural theorists who study the social construction of nature have argued that sustainability is simply another version of the “recovered garden” consisting of biodegradable industries, preservation of pristine wilderness, and social justice that finally achieves the “end drama,” a “postpatriarchal, socially just ecotopia for the postmillennial world of the twenty-first century” (Merchant 1996).

Sustainable cities require that economic, environmental, and social needs be balanced and interconnected. *Sustainability* is based on the idea that it is necessary to find the proper balance between human-made and natural environments, the “warp and woof that make up the fabric of our lives” (Van der Ryn and Cowan 1995). This constitutes a new brand of environmental thinking. Under what is

sometimes branded “the new urban ecology” (Collins et al. 2000), cities are no longer viewed as necessarily detrimental but are in fact part of the solution to environmental problems.

To make these kinds of ideals relevant, planners have translated urban impacts using concepts like “carrying capacity” to promote the idea that metropolitan development should not consume resources faster than they can be renewed or more than natural systems can process. Similarly, the “*ecological footprint*” is used to measure sustainability by calculating the amount of resources consumed, postulating that sustainable development requires reduction of ecological footprints by reducing levels of human consumption that do not exceed the ability of ecosystems to provide them. This method has lately been criticized for failing to fully account for the trade-offs and benefits of compact urban form, among other things. The ecological footprint may foster human vs. nature duality because of its emphasis on establishing a causal link between cities and accelerated global ecological decline.

54.3 Sustainable Cities and Regions

The *sustainable region* and the *sustainable city* are intertwined for one obvious reason: the *sustainable city* is almost always discussed in terms of its regional context. In a planning sense, regionalism is about the pattern of human settlement (villages, towns, and cities) set in protected open space. Anyone advocating the development of self-contained units of human settlement knows that these units must be positioned geographically but also that it is necessary to think of them in terms of an integrative framework. On this point, there is little disagreement, and the idea has been operative since the regionalist perspective applied to city planning came into fruition more than 100 years ago.

Looking at the world from a regional perspective began even earlier and can be traced back as far as the eighteenth century when the natural and cultural geography of Europe was particularly suited to regional differentiation. A number of definitions evolved, ranging from a focus on human economy and cultural distribution to the identification of natural boundaries. Peter Hall (2002) points out that the idea of towns of limited population surrounded by agricultural green belts is a recurrent theme, found in the writings of Ledoux, Owen, Pemberton, Buckingham, and Kropotkin, and More, Saint-Simon, and Fourier had cities arranged within a regional complex.

Sustainability at the regional level revolves around a related idea that issues like housing, *transportation* and the environment, and the political governance of each must be treated as an interconnected, multijurisdictional whole. There is a need to balance human activity and nature by keeping settlement at the proper scale and level of self-sufficiency. The *Regional City* envisaged by contemporary regionalists (Calthorpe and Fulton 2001) conceptualizes the “emerging region” as a revitalized central city coexisting with strengthened suburbs and preserved natural areas.

In the USA, sustainable cities are a more common conception than sustainable regions, although in Europe, this is not the case. In all regions, planners seem

especially in agreement about what an *unsustainable* city looks like. There is no disputing that malls surrounded by parking lots, disconnected apartment complexes, and vast expanses of low-density detached housing – that is, sprawl – have a higher carbon footprint than attached buildings in a walkable context. The recent book *Green Metropolis* by David Owen (2009) presented the most thorough documentation of this fact to date: compact neighborhoods bring with them the intrinsic environmental, social, and economic benefits of living smaller, driving less, lowering energy costs, strengthening social connection, and fostering networks of economic interdependence. The book *Sustainable Urbanism* by Doug Farr (2008, p. 10) provided an even more explicit vision: “Sustainable urbanism is an integration of walkable and transit-served urbanism with high-performance buildings and high performance infrastructure.” Planners largely agree that the *sustainable city* is more than just green buildings and pervious pavement; it involves the design of walkable communities along with the connection to transit, food, and amenity they require.

How does urban planning support sustainable ideals more specifically? Of course, the concept of a “*sustainable city*” includes more than the physical and infrastructure qualities of built form. In particular, institutional strategies like recycling programs, local governance, and civic participation are considered important for promoting sustainable cities. And always, green building and infrastructure technologies – efficiencies in structural design, energy use, and materials, as well as green infrastructure – are an important part of the task of city building.

We can summarize the key principles that promote the *sustainable city* from an environmental point of view. Cities must (a) lower vehicle miles traveled (VMTs), limiting carbon emissions by looking for ways to reduce reliance on fossil fuels (cars) and increase reliance on clean *transportation* (e.g., bus rapid transit, light rail); (b) lower energy costs by lowering infrastructure, like highways, and utility lines, which in turn results in lower transmission loss; and (c) limit damage to natural environments by lowering impervious surfaces and runoff, compacting development, and lowering disruption of *biodiversity* and natural habitat. Sustainable industrial and energy systems, food production, and mitigation of heat-island effects are also essential. Sustainable cities also promote “green streets” that handle stormwater within their right-of-way; contain visible, green infrastructure; and maximize street trees to improve air quality, reduce temperature, and absorb stormwater. Sustainable cities support passive solar design, sustainable stormwater practices, organic architecture, the harnessing of waste heat, and the protection of *biodiversity* corridors. Local jurisdictions in the USA have been attempting to incorporate sustainability in their activities, regulations, and development approval processes, promoting eco-industrial park development, bicycle ridership programs, point systems for green architecture, or the use of sustainability indicators.

That is just the environmental side. To endure economically, the *sustainable city* needs to foster diverse economic networks of interconnected relations, a view that Jane Jacobs famously advocated in *The Death and Life of Great American Cities* (1961). The basic idea is this: “the combinations of mixtures of activities, not separate uses, are the key to successful urban places” (Montgomery 1998, p. 98).

Allan Jacobs and Donald Appleyard wrote a widely cited manifesto in which they argued that *diversity* and the integration of activities were necessary parts of “an urban fabric for an urban life.” The maximizing of “exchange possibilities,” both economic and social, is viewed as the key factor of urban quality of life (Greenberg 1995) and, now, *sustainability*. What counted for Jane Jacobs was the “everyday, ordinary performance in mixing people,” forming complex “pools of use” that would be capable of producing something greater than the sum of their parts (Jacobs 1961, pp. 164–165).

Thus, sustainable cities are tied to, or ultimately derived from, social and economic *diversity*. The book *Building Sustainable Urban Settlements* (Romaya and Rakodi 2002), for example, lists “mixed land uses” first under its set of principles for building sustainable settlements. Reduction of travel costs, and therefore energy consumption, is usually a primary motivation. The “land use–transport connection” is put forth as a counterresponse to the problem of non-diversity, that is, functional isolation (Newman and Kenworthy 1996). A mixture of land uses has been shown empirically to encourage non-automobile-based modes of travel such as walking and bicycling, which in turn are seen as having a positive impact on public health.

An economically *sustainable city* is one that fosters opportunity. In Jacobs’ words, cities, if they are diverse, “offer fertile ground for the plans of thousands of people” (Jacobs 1961, p. 14). Non-diversity offers little hope for future expansion, either in the form of personal growth or economic development. Nor are non-diverse places able to support the full range of employment required to sustain a multifunctional human settlement. *Diversity* of income and education levels means that the people crucial for service employment, including local government workers (police, fire, schoolteachers) and those employed in the stores and restaurants that cater to a local clientele, should not have to travel from outside the community to be employed there.

This brings us to the third dimension – that of social *sustainability*. As with economic sustainability, *diversity* is seen as a key variable. A socially diverse city – one that avoids differentiation of social groups into segregated housing enclaves – ensures better access to resources for all social groups, providing what is known as the “geography of opportunity” (Briggs 2005). *Diversity* builds social capital of the bridging kind by widening networks of social interaction. Where there is less social *diversity* and more segregation, there is likely to be less opportunity for the creation of these wider social networks. This could be a significant disadvantage for segregated neighborhoods and could even have the effect of prolonging unemployment.

While socially diverse neighborhoods continue to be seen as essential for broader community well-being and social equity goals, the connection to *sustainability* is also made – mixing incomes, races, and ethnicities are believed to form the basis of “authentic,” sustainable communities (Talen 2008). In addition to mixed housing type, land uses that complement each other to promote the active use of neighborhood space at different times of the day will create “complex pools of use” (Jacobs 1961), a component of natural surveillance and social

sustainability. Supporting this are findings that a mix of neighborhood public facilities plays a role in reducing crime. Studies of socially mixed neighborhoods consistently identify *urban form* as a key factor in sustaining *diversity*.

It is possible to focus more specifically on the human-built dimensions of *urban form* – streets, lots, blocks, land uses, buildings, and the patterns they create. A kind of meta-principle for sustainable *urban form* is compactness. All of the environmental principles of *sustainability* suggest or even require it. Some of this is obvious. Compactness means that there will be fewer highways, greater transit feasibility, greater opportunities for combined heat and power, and lower pumping requirements for water and sewer. Conversely, low-density development has been linked to higher infrastructure costs, increased automobile dependence, and air pollution. Density has been seen as an essential factor in maintaining walkable, pedestrian-based access to needed services and neighborhood-based facilities, as well as a vibrant and diverse quality of life (Newman and Kenworthy 2006).

Walkable access to services is an essential part of the *sustainability* equation because people living in well-serviced locations will tend to have lower carbon emissions (Ewing et al. 2008). The higher the access to opportunities like jobs and services, the lower the *transport* costs. Related to this, sustainable *urban form* is defined by the degree to which it supports the needs of pedestrians and bicyclists over car drivers. This has been motivated by a concern over the effect of the built environment on physical activity and human health. Streets that are pedestrian-oriented are believed to have an effect not only on quality of place but on the degree to which people are willing to walk. Researchers have argued that activity levels can be increased by implementing small-scale interventions in local neighborhood environments, and a whole catalog of design strategies are now used to make streets more pedestrian-oriented.

Finally, sustainable *urban form* is associated with what could be termed polycentric or multinucleated urbanism – the idea that urban development should be organized around nodes of varying sizes (see Frey 1999). Whereas sprawl tends to be spread across the landscape uniformly, sustainable *urban form* has a discernible hierarchy to it – from regional growth nodes to neighborhood centers or even block-level public spaces. At the largest scale, centers may be conceived as regionally interconnected “urban cores,” with higher intensity growth converging at *transportation* corridors. At the neighborhood level, nodes support sustainable *urban form* by providing public space around which buildings are organized.

54.4 Implementing Sustainability Goals

While there is wide agreement among urban planners on the principles outlined above, it is during implementation of these ideals that tensions about sustainability are exposed. In fact, there is plenty to debate, and those debates frame some of the most interesting aspects of *sustainability* in planning.

Debates about *sustainability* in planning are often a matter of degrees. Studies of the connection between, say, *urban form* and travel behavior, or between *urban*

form and health, may admonish planners for failing to see the full complexities involved in linking particular forms to *sustainability* outcomes, but they are unlikely to call for the wholesale reversal of the basic idea that compact, diverse, walkable urbanism supports *sustainability*. Between business-as-usual sprawl development of the “nineteen real estate product types” that define suburbia (Leinberger 2008) and urban planners seeking compact urbanism, there is a significant divide. Debates on the planning side center on how much compactness, walkability, and *diversity* of land use, not whether compactness and *diversity* are important goals. While it is entirely possible to discuss the level of social *sustainability* – in terms of equity, justice, and social capital – that might be impacted by alternative *urban forms*, these too are a matter of degrees (Ancell and Thompson-Fawcett 2008).

One source of debate is over the degree to which partial success in implementation is better or worse than doing nothing. Ancell and Thompson-Fawcett raise a legitimate question about whether intensifying parts of a city makes a city more sustainable: “if this results in diminished opportunities for lower-income groups to live in the central city, is such intensification necessary or sufficient as a basis for social *sustainability* with respect to planning for housing?” (2005, p. 427). One response is that this might be more a matter of failed implementation than of failed principle.

But what might be more elusive and interesting is not the degrees to which different sustainable *urban forms* achieve their desired purpose and not whether implementation is actually occurring – although both of these questions are critical and continue to engage researchers – but whether there are debates that are even more fundamental. There are two debates in particular: process vs. form and city vs. nature.

The first concerns the tension between flexibility and open-ended process vs. preconceived forms and concrete visions expressed as ideal models of *urban form*. In the book *The Original Green* by Steven Mouzon (2009), the argument is made that cities, to be sustainable, must embody a number of specific design principles, from walkable streets to preservation of the embodied energy of historic buildings, to design that encourages the use of public space and the civic interaction that results. But for some, the specific design qualities needed to enhance *sustainability* at this level should be more open ended and flexible – not normative.

In lieu of the normative approach of new urbanists, some argue that *sustainability* in planning amounts to managing “the continuous processes of change” (Brown 2006, p. 100). Brown notes that this was the perspective of the revered urban planner Kevin Lynch, who wrote in the book *Good City Form* (1981): “The good city is one in which the continuity of this complex ecology is maintained while progressive change is permitted. The fundamental good is the continuous development of the individual or the small group and their culture: a process of becoming more complex, more richly connected, more competent, acquiring and realizing new powers – intellectual, emotional, social, and physical” (Lynch 1981, p. 116). Planners and designers who agree with this view are against “a steady state with respect to human–environment relations” and instead devote their energies to promoting the best possible process.

Process and form are not necessarily in opposition. New urbanists would contend that both are needed, relying on the charrette process to implement their model of sustainable *urban form*. But many planners, while they would agree with the basic outlines of what a *sustainable city* looks like, are more interested in ensuring a sustainable process for getting there. Phil Berke summed up what the authors in a special issue on green communities in the *Journal of the American Planning Association* called for: “collaborative planning processes aimed at strengthening and mobilizing social networks to support green community initiatives, requirements and incentives that stimulate greener community and household behaviors, and new assessment tools for green building rating, and greenhouse gas inventory and analysis” (Berke 2008, p. 393). In other words, the focus is on process, procedure, and assessment rather than specific design ideals. Largely this entails prioritizing resident views: “it is important to avoid undertaking research with pre-conceived notions as to whether impacts of urban compaction such as smaller houses are negative or positive, and instead to let the residents speak for themselves” (Ancell & Thompson-Fawcett, 2008, p. 440). Others have argued that the *sustainable city* is being thwarted by “unresponsive bureaucratic procedures” that are ill prepared to deal with the reality that “sustainable development is political rather than analytical” and that overly pragmatic policy solutions (i.e., urban design ideals) might forever frustrate the value-laden complexity of *sustainability* (Batty 2006, p. 38).

In the architecture field, models of sustainable *urban form* that appear to be universalist (such as those of the new urbanists) are rejected. Architects are especially “skeptical of the assumption that a single approach, model, or list of best practices can be universally applied,” arguing instead for a “much needed transdisciplinary conversation to emphasize the long-term consequences of our actions, not their ideological or disciplinary purity” (*Pragmatic Sustainability: Theoretical and Practical Tools* (<http://www.routledge.com/books/search.aspx>)). Reviewing why one city is better able to develop sustainably than another rests on “particular dispositions toward politics, nature, and technology,” not “a single abstract model” (Moore 2007).

An even more fundamental debate concerns the relationship between cities and nature, which can be described as the “human vs. nature duality.” There is a long history to this in urban planning, and the focus on *sustainability* has not escaped it. It comes from the regionalism of early-twentieth-century botanist Patrick Geddes, who viewed metropolitan development as dependent upon knowledge of the large-scale, regional complexities of the landscape and the human response to that landscape. However, early regionalists believed no synthesis between *existing* metropolitan development and nature was possible. This imbalance, which was explicitly outlined by MacKaye (1928) in *The New Exploration*, came to epitomize the view that large metropolitan areas were the antithesis of environmental conservation.

Historian William Cronon explored the phenomenon of separating human and natural worlds in the book *Uncommon Ground: Rethinking the Human Place in Nature* (1996). He argued that wilderness, the “ideological underpinning” of the environmentalist movement, is a highly problematic concept because it is viewed

as something wholly separate from ourselves. Even the opening line of the Union of Concerned Scientists' *Warning to Humanity* (1992) included the premise of separation; it begins: "Human beings and the natural world are on a collision course."

What may be the most lucid example of human/nature duality in planning is the way in which the "greening" of human places is interpreted as something unilaterally positive for the environment, regardless of broader impacts. There may be a failure to recognize that metropolitan development patterns that appear "natural" in the suburban landscape actually disrupt natural systems. In fact, maintaining green spaces may be harmful both in direct ways (through soil compaction, irrigation, and the need for chemical treatment) and in indirect ways – increasing atmospheric pollution through increased automobile use caused by spreading out the urban pattern. In short, interweaving green spaces through human settlement may sometimes be more harmful than not when viewed at a larger scale. Somewhat ironically, the most environmentally sound pattern of human settlement – in some cases – may be the one with lower rather than higher levels of green space.

This tension between cities and nature has been identified by Godschalk and others as the "green cities conflict." It is essentially a conflict over the degree to which natural vs. human connectivity is to be prioritized. New Urbanism has been criticized for failing to accommodate more environmental sensitivity (Berke 2002), and this often boils down to their focus on maintaining urban connectivity. But new urbanists counter that environmental regulations may inadvertently thwart compact urban development, including suburban retrofits. They point to a potential problem with Low-Impact Development, which is attempting to replace the old stormwater system approach of "pave, pipe, and dump engineering" with something more eco-friendly. The old system resulted in high runoff rate, volume, and pollutant loads and needed to be changed. But new stormwater regulations, currently being advocated by the US Environmental Protection Agency (EPA), might actually incentivize sprawl and reduce retrofits by making each site "emulate the natural hydrologic conditions of the site." Since this becomes much easier in greenfield sites but is onerous on redeveloped sites, the urbanization of existing places may ultimately be avoided.

An alternative to the Low-Impact Development approach has been proposed in the *Light Imprint Handbook: Integrating Sustainability and Community Design* (<http://www.lightimprint.org>). The book lays out environmentally friendly stormwater management and includes 60 techniques for "paving streets and walkways, channeling and storing water, and filtering surface runoff before release into the aquifer." It is described as "an intrinsically green design strategy" that not only sustains compact urban places but also "respects site terrain." The approach is based on the idea of an urban-to-rural transect as a way of maintaining the proper interconnections among urban elements – a balanced mix of landscape, building type, and streetscape, for example.

How should compact urban development integrate with green infrastructure practice? Projects that are greenfield rather than infill, disconnected from existing

infrastructure, and not particularly concerned with stormwater runoff and the restoration of wetlands are problematic (Berke et al. 2003). But it is also true that environmental “Best Management Practices” and “Low-Impact Development” can result in a lack of urban connectivity, undermining the ability to develop walkable, diverse, compact places – sustainable *urban form*.

54.5 Conclusions

There is an interesting overlap between the two debates discussed above: the need to balance process vs. form and the need to integrate city vs. nature may ultimately converge in our approach to sustainable urban planning. For example, it could be argued that creating visually explicit models of future development *and* providing an inclusive process are both needed to help resolve the human vs. nature duality problem. Perhaps development that is represented tangibly (as compact urban neighborhoods) can help overcome human/nature duality by helping people visualize how development that meets human needs can also protect natural areas. Recognizing that urban development is not a zero-sum game with trade-offs between social and environmental goods, normative visions of development could be used to help illustrate the possibilities. And an inclusive process that allows flexibility and the exploration of alternative proposals is needed not only to ensure that development actually addresses human/nature integration but that it does so in a way that makes sense to people.

The definition of the *sustainable city* and region is, in principle, resolved. The question is how to get there: via a reliance on the right process that guides city building toward a more sustainable outcome, or via a stronger articulation of what the *sustainable city* and region is supposed to be, or via an urban development approach that prioritizes natural systems or one that allows natural systems to be trumped in some cases in order to promote urban connectivity and compactness? City and regional planners need to find a balance between visualized ideals and inclusive process and between the unequivocal protection of nature and the corresponding human claim to land development. In both cases, there is a need to bring the language of integration into sharper focus. In sustainability, actions are supposed to balance natural, economic, and social concerns. *Sustainability* challenges us to make every decision supportive, and integrative, of each realm.

References

- Ancell S, Thompson-Fawcett M (2008) The social sustainability of medium density housing: a conceptual model and christchurch case study. *Hous Stud* 23(3):423–441
- Batty SE (2006) Planning for sustainable development in Britain: a pragmatic approach. *Town Plan Rev* 77(1):29–40
- Berke P (2002) Does sustainable development offer a new direction for planning? Challenges for the twenty first century. *J Plan Lit* 17(1):22–36

- Berke P, Manta-Conroy M (2000) Are we planning for sustainable development? An evaluation of 30 comprehensive plans. *J Am Plan Assoc* 66(1):21–34
- Berke P, McDonald J, White N, Holmes M, Oury K, Ryznar R (2003) Greening development for watershed protection: does new urbanism make a difference? *J Am Plan Assoc* 69(4):397–413
- Berke PR (2008) The evolution of green community planning, scholarship, and practice. *J Am Plan Assoc* 74(4):393–408.
- Brown DF (2006) Back to basics: the influence of sustainable development on urban planning with special reference to montreal. *Can J Urban Res* 15(1 Suppl):99–117
- Calthorpe P, Fulton W (2001) The regional city: planning for the end of Sprawl. Island Press, Washington, DC
- Campbell S (1996) Green cities, growing cities, just cities? Urban planning and the contradictions of sustainable development. *J Am Plan Assoc* 62(3):296–312
- Collins J et al (2000) The new urban ecology. *Am Sci* 88:5
- Cronon W (ed) (1996) Uncommon ground: rethinking the human place in nature. W.W. Norton and Co., New York
- Daniels TL (2009) A trail across time: American environmental planning from city beautiful to sustainability. *J Am Plan Assoc* 75(2):178–193
- De Souza Briggs X (ed) (2005) The geography of opportunity. race and housing choice in metropolitan America. Brookings Institution Press, Washington, DC
- Ewing R, Keith B, Steve W, Jerry W, and Don C (2008) Growing cooler: the Evidence on urban development and climate change. Washington, DC: Urban Land Institute
- Farr D (2008) Sustainable urbanism: urban design with nature. Wiley, Hoboken
- Frey H (1999) Designing the city: towards a more sustainable urban form. Taylor and Francis, London
- Greenberg M (1995) The poetics of cities: designing neighborhoods that work. Ohio State University Press, Columbus
- Hall P (2002) Cities of tomorrow: an intellectual history of urban planning and design in the twentieth century, 3rd edn. Basil Blackwell, Oxford
- Jacobs J (1961) The death and life of great American cities. Vintage Books, New York
- Leinberger C (2008) The option of urbanism. Island Press, Washington, DC
- Lynch K (1981) Good city form. MIT Press, Cambridge
- MacKaye B (1928) The new exploration: a philosophy of regional planning. Harcourt, Brace and Co., New York
- Merchant C (1996) Reinventing Eden: western culture as a recovery narrative. In: Cronon W (ed) Uncommon ground: rethinking the human place in nature. W.W. Norton, New York, pp 132–170
- Montgomery J (1998) Making a city: urbanity, vitality and urban design. *J Urban Des* 3(1):93–116 (p. 98)
- Moore SA (2007) Alternative routes to the sustainable city: Austin, Curitiba and Frankfurt. Rowman and Littlefield, Lanham
- Mouzon SA (2009) The original green: unlocking the mystery of true sustainability. The New Urban Guild Foundation, Miami
- Newman PWG, Kenworthy JR (1996) The land use–transport connection: an overview. *Land Use Policy* 13(1):1–22
- Newman PWG, Kenworthy JR (2006) Urban design to reduce automobile dependence. *Opolis: Int J Suburb Metrop Stud* 2(1):35–52
- Owen D (2009) Green metropolis: why living smaller, living closer, and driving less are the keys to sustainability. Riverhead Hardcover, New York
- Romaya S, Rakodi C (2002) Building sustainable urban settlements: approaches and case studies in the developing world. ITDG Publishing, London
- Talen E (2008) Design for diversity. exploring socially mixed neighborhoods. Elsevier, London
- Union of Concerned Scientists (1992) World scientists' warning to humanity. Union of Concerned Scientists, Cambridge, MA. Available at: <http://www.ucsusa.org/>

- Van der Ryn S, Calthorpe P (1991) Sustainable communities: a new design synthesis for cities, suburbs and towns. Sierra Club Books, San Francisco
- Van der Ryn S, Cowan S (1995) Ecological design. Island Press, Washington, DC
- World Commission on Environment and Development (the Brundtland Commission) (1987) Our common future. Oxford University Press, Oxford

Jill L. Findeis and Shadayen Pervez

Contents

55.1	Introduction	1086
55.2	Century of Dramatic Population Growth and Population-Environment Theories	1088
55.3	Regional Differentials	1092
55.4	Micro Perspective on Population Decisions and Community Resilience and High-Level Effects	1095
55.5	Conclusions	1101
	References	1102

Abstract

The impact of human population growth on the environment represents the major challenge of our time. This chapter reviews demographic change over the last century, set in historical context, and different perspectives on population-environment interactions. Differences in population growth rates and demographic change across space are explored, followed by perspectives on the population-environment nexus at multiple scales with a particular focus on those contexts where impacts are likely to be the very greatest on humankind. The alignment of individual and higher-level actions resulting in environmental impacts and the negative force of the impacts of actions are important, to signal the need to change behaviors. Interrelationships are shown to be highly complex. It is argued that multidisciplinary efforts to tackle complexity and to focus on resilience at multiple scales are critically needed, with the importance of multidisciplinary regional science thought being underscored. The question is raised, however, whether

J.L. Findeis (✉) • S. Pervez

Division of Applied Social Sciences, University of Missouri-Columbia, Population Research Institute, Pennsylvania State University, Columbia, MO, USA
e-mail: findeisj@missouri.edu; fa2@psu.edu; shadayen@gmail.com

these efforts will be coordinated well enough across multiple scales and with multiple disciplines and publics to avoid what could be catastrophic impacts. These are most likely to occur at local and regional scales where population growth rates are high, natural environments already vulnerable, and resilience limited.

55.1 Introduction

Vitousek et al. (1997), among many other scholars, convincingly argue that the impact of human population growth on the environment represents the major challenge of our time. Increasingly, humans dominate the landscape, driven by population growth, its distribution and affluence. Population growth rates over the last century are unprecedented, stemming from multiple factors including higher survival rates of the young and old, albeit coupled with a recently falling birth rate (UNFPA 2011). World population now exceeds seven billion. Population Reference Bureau (PRB 2010) projections to 2050 indicate a doubling of the world's population from 2010 levels in the least-developed countries (i.e., 2.0 times more), 1.6 times more in the less-developed countries excluding China, 1.4 times more in the less-developed countries with China included, and 1.1 times more in more-developed countries. Simultaneously, economic growth and development has been attained in many regions of the world previously considered – even 50 years ago – as almost endlessly trapped in poverty.

The physical environment, which historically kept human population growth in check, has been altered to the point that many question whether and what individual and collective human actions are needed to exert the impact that the environment previously – and harshly – did. The tables are turned; human population growth, expected to continue over the next century, threatens the natural environment even as overall rates of population growth are projected to decline. Caveats are in order. The long-term perspective is important, as documented by historian J. R. McNeill (2006). And the epidemiological and agricultural transitions that contributed to higher survival rates are known to be reversible. However, Campbell (2007) challenges us when she observes that recent literature across multiple disciplines remains amazingly silent on the population growth issue. In part this is a result of conflict over the role of family planning. What is clear is that the complexity of the population-environment nexus and underlying mechanisms are major challenges for scientists across a wide spectrum of disciplines.

Ehrlich and Holdren (1971) summarized the complexity of the population-environment nexus in the famous $I = PAT$ equation. Here, the aggregate environmental impact (I) is the product of the size of the population (P), per capita consumption (A for affluence level), and the environmental impact of a unit of consumption determined by level of technology T . The basic assumption is that environmental impacts, whether level of pollution or depletion of the natural resource base, are functions of the total demand for goods and services in the economy. Technological improvement implies discovery of ways to substitute natural resources among one another and to substitute ideas and manufactured

capital for natural resources. This enables higher levels of production and consumption even without making higher demands on the natural resource base. IPAT is evidently too simplistic: it does not account for the interactions among the variables. For instance, technology can be more efficient as the level of affluence increases. Another serious shortcoming is that it specifies impact as a linear relationship ignoring the very important threshold effects. But despite these simplifications, IPAT points out the ruthless logic of the well-known Malthusian framework: the ever-growing population and level of affluence will continue to have a greater toll on the earth's natural resource base and ecosystem. To what extent technological progress can stretch the eventual limit of the earth's carrying capacity is a source of continuing debate, but the simple IPAT equation underscores the apparent problem of assuming an infinitely elastic limit.

Viewed thus from a Malthusian perspective, IPAT points to an impending "population problem." Scientists from different disciplines have different ideas on why the problem exists. For example, for economists, one challenge is to find out why the problem at the aggregate level cannot be solved at the level of individual rational actors. Given the constraint on the natural resource base, rational actors should adjust their reproductive and consumption behaviors so that the optimum levels of population and consumption can be perpetually maintained. Of course, this does not happen in reality in part because the private and social benefits and costs of reproduction are not the same. Reproduction generates crowding externalities for others, which the reproducing individuals acting on their own are not expected to take into account. One likely source of these crowding externalities is the obvious finiteness of space. Thus, crowding externalities provide population-environment interaction a spatial perspective, making it of paramount interest to every regional scientist.

The overall population growth trend is coupled with an apparent unevenness in population growth rates across regions: in Japan and the EU, growth rates are negative, while numbers across the developing world continue to climb. The totals projected in the literature for some countries are really astounding. As documented by Campbell (2007), by 2050, Nigeria expects to grow from 27 to 131 million, Niger from 14 to 50 million, and India "by a net million every three weeks, nearly all of this growth in the lowest income regions of the country" (Campbell 2007, pp. 237–238). This unevenness has already led to human suffering in place. It has led to forces pushing internal migration to find food and water. Transnational and transregional migration is common. While the pull of better opportunities elsewhere is a major incentive, higher than average population growth rates in already resource-poor regions of the world set in motion forces that also push population out of homelands, putting additional pressure on regions where resources are more abundant and/or native populations are declining. The developed countries are not off the hook in terms of the impacts of their *own* behaviors on the natural environment. Over the past half century at least, the trend toward higher levels of consumption is an important driver of environmental degradation. *The ability to have more* puts stress on the natural environment not only in the developed world but also in developing countries and regions, often important suppliers of specific natural resources. Further, imitation within developing countries can be an important driver of innovation.

Recent demographic trends raise a host of challenges to scientists and to the world's public: how to efficiently and equitably feed the human population concentrated in urban centers or the elderly who are becoming a larger proportion of the rural population; how to maintain human health and well-being in these places in the long run; how to sustain the natural resource base in populated spaces; and how to design appropriate policies to balance social, economic, and environmental goals. Addressing these challenges requires understanding the complexity of the human system and larger ecosystem, which underlie the population-induced transition process. A study of recent debates in the literature on the interrelationships among human population growth, economic growth and development, and environmental impacts will convince most that the underlying system is complex, requiring the cooperative problem-solving that some believe humans may be genetically programmed to do best.

This chapter reviews selected works from multiple literatures contributing to disciplinary, multidisciplinary, interdisciplinary, and transdisciplinary efforts to understand and solve the complex challenges posed by the population-environment problem. After reviewing the historical context and documenting population growth and distribution trends, literature at different scales – micro, meso, and macro – is reviewed. The underlying question to be answered by future researchers is the following: Can a balance between humans and the environment be struck to attain a sustainable state? And, importantly, how can regional scientists contribute to the process of solving the population-environment challenge?

55.2 Century of Dramatic Population Growth and Population-Environment Theories

The “Malthusian trap” has fortunately not (yet) been realized. In *An Essay on the Principle of Population as It Affects the Future Improvement of Society*, Malthus (1798) argued that the rate of growth of population would outstrip growth in food production, the latter being constrained by increasingly scarce land resources. He believed that human fertility could not be curtailed, that is, humans would reproduce without restraint, although Malthus later softened his stance. Birdsall (1988) observes that while population growth happened in the eighteenth and nineteenth centuries, growth was slow, “seldom exceeding one percent a year” (Birdsall 1988, p. 478). From 1750 to the twentieth century, the overall population growth rate averaged 0.5 % per year, with rates higher in what are now the more-developed countries. While couples likely controlled their own fertility to a greater extent than we perceive, “environmental conditions” (notably disease) limited growth – to be young puts you at peril, and relatively fewer lived to be “elderly” as we know it today.

It is important to emphasize at this juncture that the Malthusian emphasis on the “direct race” between population growth and food growth fails to uncover the underlying forces behind the mechanism at work. Production after all depends on the level of employment, not the level of population. But there is not a one-to-one mapping either between population growth and employment growth or between employment growth and food production growth. Over the last two and a half

centuries, changes in fertility and life expectancy obviously have changed the labor force participation rate or the employment level.

Technological change also has boosted agricultural productivity, yielding higher levels of food production for the same level of employment. Higher aggregate income may not result in proportionate food growth. Further, growth in the food sector crucially depends on the distribution of income. Technological change is again instrumental in the determination of factor rewards and income distribution. Thus, the determining factor behind the Malthusian thesis is the extent to which the forces of technological change affect the key relationship between population and food growth.

Extended Life Spans, the Epidemiological Transition and Voices of Concern.

Economic growth and development and technology, in particular, extended life spans in the years following World War II. The “epidemiological transition” was built on advances in public health through disease prevention (e.g., immunization, improved sanitation) and effective forms of disease control (e.g., antibiotics). Technological advances in agriculture contributed to higher yields and greater food security in many of the world’s regions and to arguably better nutritional status. Significantly lower mortality rates emerged, which when coupled with continued high fertility rates, resulted in population growth rates of 2.4 % per year in the 1960s (Birdsall 1988). High population growth rates were especially concentrated in the less-developed countries; Europe had less than a 1 % rate of growth and North America exceeded “1.5 % only briefly” (Birdsall 1988, p. 479). Japan’s rates also were low. However, the higher than replacement rates documented in almost every country of the world raised significant concern. As early as 1953, the United Nations published *The Determinants and Consequences of Population Trends*, and Coale and Hoover (1958) followed with *Population Growth and Economic Development in Low-Income Countries*. Human population growth coupled with economic growth and development over this time period began to raise questions about the ability of the earth’s natural resources – the so-called Malthusian “flower pot” – to sustain the unprecedented growth.

The voice of concern became even louder in the 1960s and early 1970s, when academics from multiple disciplines focused energy and public debate on two alarming trends – emergence of signs of environmental degradation at multiple scales worldwide tied to ever more population coupled with greater affluence in some regions. *The Population Bomb* (1968) credited to Paul Erlich (and Anne Erlich); *The Limits to Growth* commissioned by the Club of Rome and authored by Donella Meadows, Dennis Meadows, Jorgen Randers, and William Behrens III; and Rachel Carson’s earlier (1962) *Silent Spring*, among other published works, raised public awareness and alarm. Publication of the National Academy of Sciences’ *The Growth of World Population* in 1963 further contributed to understanding by the public and academic community. *The Limits to Growth* and *World Dynamics* by Jay Forrester are exercises in system dynamics and cutting-edge computer simulation that gave those works an aura of scientific precision and increased respectability.

The simultaneous growth in population and affluence triggered the Neo-Malthusian alarm. The essential logic rests on the imperfect substitutability

between natural and man-made resources. Given imperfect substitutability, if land and natural resources remain fixed, growth in labor and physical capital cannot sustain output growth forever; it is only time before diminishing marginal returns to labor set in and per capita production and consumption begin declining. *The Limits to Growth* emphasizes this point by asserting that limits to arable land will soon be reached and per capita consumption will decline, leading to famine. Even if a food crisis is avoided, the demand for industrial output will exhaust the earth's mineral resources; the growing pollution resulting from industrial output will generate pollution levels beyond the earth's exhaustive capacity, eventually leading to catastrophic collapse.

Neoclassical and Cornucopian Thought. In a devastating but pointed criticism, Nordhaus (1973) dismissed the models in *World Dynamics* and *The Limits to Growth* as devoid of empirical content. The predictions of these models are highly sensitive to specifications. He objected that the models that had been constructed were based (merely) on subjective assumptions and were not reconciled to the real world. Thus, if the assumptions did not hold, the results indicated by the models would not hold either.

The dire Neo-Malthusian predictions stem from simplifying assumptions that can indeed be challenged. Natural resources may not be fixed: new natural resources are being discovered through relentless exploration, although even here limits to growth could eventually become binding. But within a well-functioning market system, the prices of scarcer resources will rise, providing incentives to substitute for them. The work of Ester Boserup demonstrated that the post-industrial revolution in agricultural production occurred side by side with population growth as a result of the greater substitution of capital and labor for natural resources. The Boserup hypothesis is thus the widely known claim that population growth triggers production increases through intensification – greater use of capital and labor instead of land. From this perspective, the key barriers to reaching an optimal production level are either market failures (such as insecure property rights, pervasive externalities) or policy distortions or both.

The discovery of ways to substitute abundant resources for scarcer ones is itself an important form of technological improvement. Technological innovation responds to incentives in the market. For example, Popp (2002) provides systematic evidence using patent data from 1970 to 1994, documenting the impact of energy prices on patents for energy-saving innovations. The propensity of technological innovation to respond to incentives had been emphasized by Julian Simon in *The Ultimate Resource* (1981) and *Theory of Population and Economic Growth* (1986), as well as other works. Simon argues that since human ingenuity to discover new ideas is the ultimate resource, population growth is the greatest boon. More people mean more ideas, that is, a greater number ways to substitute and circumvent resource constraints. Larger population also means a greater size of the market, greater specialization, and greater per capita productivity. Thus, as opposed to the future of the Malthusians, Simon's ideas prognosticate a rosy future of material abundance – hence the name cornucopian theory.

The relatively recent development of Paul Romer's New Growth Theories provides impetus to the cornucopian school. Investment in research and development

yields increasing returns to scale because the benefits of ideas do not diminish through sharing. Yet the logical corollary to cornucopian thought is the claim that infinite growth in output makes no more than a finite demand on the natural resource base. This claim is famously refuted by stating that no degree of production intensification in a flower pot can potentially grow enough to feed the entire world.

Quelling some concern, population growth rates in almost all developed countries declined to replacement rates in the 1970s, although rates in the developing world remained high. While in part in response to growing public discourse on *The Population Bomb*, *The Limits to Growth*, and *Silent Spring* (among others), more importantly (and pragmatically) this decline was in response to the advent of widespread availability of more effective fertility control and the return of women to paid jobs. This occurred regardless of the income level of the country; Birdsall (1988) makes the compelling case that while countries with higher incomes tend to have lower rates of fertility and mortality, many countries with individual and household low incomes have achieved lower fertility and mortality rates too. Lower rates were achieved through advances in women's status and greater access to modern fertility control, education, and health services.

Two forces diffusing the concern at this stage are the following: (i) the second demographic transition, the gradual decline of fertility approaching replacement level, and (ii) the environmental Kuznets curve (EKC), the empirical observation that environmental quality *appears to* initially deteriorate with economic growth but then improves as income growth continues. These two phenomena working together mitigated some of the alarm of impending catastrophe and modified public and academic discourse on the population-environment problem.

Second Demographic Transition and the EKC. For most of human history, the rise of per capita income had a positive effect on population growth. Population growth in turn caused diminishing marginal labor productivity to set in, and the Malthusian check eventually reduced per capita income to near subsistence levels. Low mortality with continued high fertility, a phenomenon called the first demographic transition, similarly could have a diluting effect on income per capita. However, this dilution of per capita income was counteracted by the acceleration in technological progress and capital accumulation.

An important recent development has been the crucial role of human capital. Galor (2005) argues that further acceleration in the rate of technological progress increased the demand for human capital in the second phase of the industrial revolution, inducing parents to invest in their children's human capital. The rise in life expectancy also increased the rate of return to investment in human capital. While human capital investment increased productivity and the opportunity cost of time, advances in household production technology, the introduction of (more) fertility control technologies, and changes in gender norms and the institution of marriage gave women greater control over reproduction. The net effect has been Gary Becker's quantity-quality trade-off in fertility choice. Thus, the second demographic transition of the post-Malthusian epoch began where sustained economic growth coincided with the simultaneous decline in fertility rates.

Further, it was observed that since the 1990s, the relationship between environmental degradation and per capita income has exhibited an inverted-U shape (Grossman and Krueger 1995). Analogous to the Kuznets curve between per capita income and income inequality, this relationship is aptly named the environmental Kuznets curve (EKC). The primary reason behind the EKC is perhaps structural change. Early stages of economic growth are driven by industrialization – the growth in manufacturing giving rise to increases in the rate of pollution. Further, economic growth increases the share of the service sector, which is less harmful to the environment. Also, the income effect on demand for environmental quality is nonlinear – as per capita income crosses a certain threshold – consumers demand environmental quality at a much higher rate. This creates the incentive for even the manufacturing sector to employ cleaner technology. Empirically, the EKC relationship has been estimated, for example, with GEMS' (Global Environmental Monitoring Systems) emissions data as measures of environmental degradation. Recently, the EKC relationship has been questioned by Carson (2010) who provides empirical evidence to the contrary.

Recent Challenges. Many have challenged the cornucopian school's optimism and again raise significant concern over lasting and irreversible environmental damage and societal and ecosystem collapse (see contributions of Jared Diamond). The academic and public dialogue is heated but healthy. Unevenness of where population growth is taking place is creating very substantial gaps among the world's major regions; serious gaps in food security, access to water, and access to energy resources are well documented. At the same time, Nobelist Elinor Ostrom's well-known research shows that local systems can adapt positively to change, even for shared (common) resources; that is, *adaptation to preserve the local environment is possible even in light of population growth*. Whether institutions of governance at higher levels can reduce transboundary externalities (e.g., the “borderless” problems stemming from SO₂ emissions) or can be designed to do so remain critical concerns. These issues are certainly receiving much attention.

Meanwhile, global population has continued to expand and expand rapidly, although the rate of growth is finally declining. Projections to 2100 by IIASA (International Institute for Applied Systems Analysis) indicate that maximum global population will be attained in this century. As argued convincingly by Scherbov et al. (2011), uncertainty, in part stemming from “unreliable statistical information” from many areas of the world but also the inherent uncertainty of future fertility, mortality, and migration rates, casts doubt on when overall and region-specific maximums will be attained. However, they conclude that “there is little uncertainty that world population will peak and start to decline before the end of the century” (Scherbov et al. 2011, p. 575).

55.3 Regional Differentials

All countries in the world are undergoing the demographic transition (Hugo 2011). Very significant regional differentials in population growth rates exist, creating

regional variations in human and environmental conditions and also in knowledge/perceptions of the extent and character of environmental degradation. Affluence varies widely by region and within regions. The well-known contributions of Paul Krugman and others have contributed key insights into why differentials exist and persist.

Almost all population growth over the next 20 years is projected to take place in Asia and Africa and to a lesser extent in Latin America. Asia's population is estimated to be 4.2 billion, expected to peak at a projected 5.2 billion around mid-century (2052) and decline thereafter (NIC 2008; PRB 2010). Africa, having had relatively modest population growth so far, is projected to experience strong population growth until at least 2,100. This will result in a much larger population living on the African continent. In Europe, North America, Oceania, Latin America, and the Caribbean, population densities are lower and projected to be maintained at lower levels. Roughly, 1.7 billion of the world's population now lives in these regions.

Lower growth rates in the West will create larger differentials between the West and both Asia and Africa. In 1980, 24 % of the world's population lived in the West. By 2009, 18 % did; projections to 2025 indicate that this percentage will decline further to 16 % (NIC 2008). Europe and East Asia face declines in their total and working-age populations, as a result of continuous below-replacement-level rates (Hugo 2011). Western countries with relatively high in-migration rates (e.g., USA, Canada, Australia) are expected to experience population growth through immigration (NIC 2008).

UNFPA (2011) statistics indicate that 43 % of the world's population is now under 25 years of age. Regional differences in population growth rates are due, at least partially, to the large percentage of the world's population in the childbearing years and living in low-income countries. At the other end of the age spectrum, in 2011 there were 893 million people – or nearly 1 billion people – over the age of 60 years; by 2050, this figure is expected to be 2.4 billion (UNFPA 2011). Population composition could differentially influence environmental outcomes (positive or negative), due to variations in propensities for different age cohorts. For example, young adults have a higher propensity to migrate than older ones (Hunter 2000).

Worldwide, convergence in the average length of life is documented (see Edwards 2011), although country exceptions exist. The challenge of HIV/AIDS clearly has had an impact on life spans. Edwards (2011) reports that overall life expectancy in the developing world has grown by an *additional* 0.24 years annually as compared to the wealthier countries, more than doubling the rate in the developed world (Edwards 2011, p. 499). Research on within-country variation in life expectancy at birth and other survival measures underscores the underlying variability over time. Interestingly, convergence in expected length of life at birth is in contrast to divergence in income per capita (see studies referenced in Edwards 2011).

According to NIC (2008), “The ‘oldest’ countries – those in which people under age 30 form less than one-third of the population – will mark a band across the northern edge of the world map. . . . the ‘youngest’ countries, where the under-30 group represent 60 % of the population or more, will nearly all be located in Sub-Saharan Africa” (NIC 2008, p. 19). Population cohorts “in the middle” are challenged to provide for this young and rapidly growing population. Both young and

old population segments pose challenges to social and economic systems and to the working-age population challenged to provide for them. However, the impact on the environment is under debate; Dietz et al. (2007) report that while they find that population levels and affluence are drivers of environmental degradation, population age structure appears to have little effect.

Finally, structures and preferences of household units have changed, in some instances putting pressure on ecosystems. For example, University of Michigan researcher Jianguo “Jack” Liu showed that in rural China, the traditional extended household structure – with multiple families living together in the same house – is now a less-preferred lifestyle. Thus, more houses are built to accommodate the greater number of families preferring to live on their own. The research shows that this demographic change pressures Panda habitat. This represents just one of many possible examples.

Correlations Among Multiple Indicators. Overlaying maps of regional indicators of population density, population growth rates, levels of economic development, and conditions of the physical environment raises as many questions as this exercise answers. Population density and level of income fail to show a clear relationship across the earth’s regions: some regions are characterized by high population densities and are poor, whereas others are dense and wealthy. Some sparsely populated regions are rich and others poor. However, studies show that the highest population growth rates tend to occur where economic growth has been least likely (Gallup et al. 1999).

Resources in the less- or least-developed regions will be under intensified pressure as these regions cope with a larger share of the world’s population. That current population growth rates are particularly high in tropical regions is a not surprising result given recent advances in disease control and rapid adoption and diffusion of immunization, antibiotics, and other health-related technologies across the tropics. Historically, higher disease burdens, and especially a broad range of infectious diseases, have been typical. Also in the tropics, food security and agricultural productivity have been lower, given old soils (such as the soils of East Africa), human-degraded soils, higher insect pest burdens in agriculture, and high rainfall variability, among other factors.

Gallup et al. (1999) observe the following when comparing worldwide population densities, population growth rates, economic development, and underlying geography:

1. Tropical zones are “not conducive to [economic] growth,” due to presence of malaria and other diseases (Gallup et al. 1999, p. 204). But prevalence of malaria is positively correlated with population density (Gallup et al. 1999, p. 209).
2. High population density appears to be positive to (economic) growth in coastal regions but “inimical to growth in the hinterland” (Gallup et al. 1999, p. 204).
3. Geography and policy “matter,” but good policy will not be enough to counter geographical disadvantage (Gallup et al. 1999, p. 204).

In short, some of what are considered the most vulnerable places on earth are inhabited by populations with high population growth rates and lower chances of developing economically for reasons largely stemming from their natural

environment and geography. McNeill's (2006) assessment of challenges related to land use, water use, and air pollution clearly points to the pervasive issue of the distribution of resources, regional shortages, and implications for the physical environment. The upward pressure of strong population growth against an already weak natural resource base encourages a variety of responses, including the potential for multiple responses: scavenging and other survival behaviors, out-migration (Zelinsky 1971), evolution toward greater agricultural intensification, urbanization, reductions in fertility (see work of Graeme Hugo), and an array of other adaptive behaviors. Human mobility or migration in response to population change is recognized as a complex phenomenon, not easily explained. But differentials between more-developed and less-developed regions likely contribute to both "push" and "pull" incentives for internal and international migration flows.

The following section concentrates on understanding theory, and adjustments and impacts at the micro level (individuals, households, communities) and macro/regional (meso) levels. The micro perspective is covered first because population-related decisions are typically made at this level and can have local repercussions. The micro is then linked to the macro/regional (meso) perspective since environmental impacts importantly play out at these higher scales; a few examples include impacts of international and rural-urban migration, exurban land use, and the associated problem of sprawl and problems stemming from transboundary pollution. The focus of Sect. 55.4 is on less- and least-developed places because these places are among the most likely to be widely impacted by population-environment interactions. Such places also face the greatest challenges, solving the environment problem through technological change.

55.4 Micro Perspective on Population Decisions and Community Resilience and High-Level Effects

Micro Perspective. A new micro perspective has emerged in population-environment research, especially related to less- and least-developed countries. Three aspects deserve emphasis. First, a primary feature of this perspective is the focus on household-level population dynamics and relationships to environmental change. A second feature is the mediating variables approach: analysis and identification of variables (e.g., poverty, government policies, cultural norms, market or nonmarket institutions) through which population dynamics affect environmental change (Hunter 2000). Mediating variables often reinforce or even reverse the role of population dynamics in environmental degradation or enhancement. Finally, it is recognized that there is more to population dynamics than population size and growth. New micro research has gone beyond attribution of environmental degradation to increases in population, to seeking to understand how other population variables – for example, household size, age, sex composition, and migration – affect and in turn are affected by environmental change.

At its core, the micro perspective is guided by the belief that observed patterns and trends of population dynamics and associated environmental change can be

mapped to the individual household unit which makes the actual decisions about production, consumption, and reproduction. Since the pioneering work of Gary Becker, researchers have sought to identify the determinants of fertility by focusing on the reproductive decisions of individual households. Demographic research in the following half a century has uncovered that apart from the techniques available emphasized by the works of Richard Easterlin, household demand for children strongly influences fertility behavior. Demand for children in turn depends on determinants: income level, parental education, rules of inheritance, and other institutional factors, along with a host of other mediating variables.

However, a major difficulty with the focus on household as a decision-making unit is that households in reality are not monolithic. Questioning the so-called “unitary” view of the household making choices to maximize its welfare Dasgupta (2003), among many others, points to ample evidence of gender inequities within poor households; these affect allocation of education, food, healthcare, and other household resources. The unitary household model effectively ignores an important fact: if all benefits and costs are not borne by the decision maker, decisions may not be optimal. Women bear the disproportionate share of the cost of having children including the cost of pregnancy, breastfeeding, daily care, and risk of maternal mortality. Yet, in many traditional societies, men are primary decision makers as to the desired number of children. *Thus, the problem of externalities – the situation when the decision maker does not bear all the costs and benefits of the decision – prevents the optimal allocation even inside the household.*

Outside the household, mediating factors may create population-environment externalities by driving a wedge between the benefits and costs borne by the decision maker. At times, these variables can create a feedback loop that contributes to a downward spiral of resource depletion, poverty, and high fertility. The self-reinforcing patterns of sustained high fertility in the face of declining environmental resources create “vicious cycles.” The key to understanding vicious cycle in the population-poverty interaction is the nonlinear response to the change in decision variables. Poor households depend on their own unskilled household labor, low-productive agricultural land held as private property, and also surrounding environmental resources available as common property. There are thresholds in the use of each of these endowments – the point where the response to the decision variable changes nonlinearly.

Vicious Cycles, from Nutritional Status to Common Property Resources.

Partha Dasgupta outlined a pioneering vicious cycle model (see Dasgupta (2003) and references therein) where nutritional status has a critical threshold in terms of capacity to work. Below that nutritional status, labor is not productive enough to grow enough food to achieve better nutrition. Thus, the poor nutrition-low productivity state is self-sustaining. Above that threshold, a good nutrition-high productivity state is similarly self-reinforcing. For household private property, there can be a critical threshold of a productive asset (e.g., agricultural land), below which similar poverty traps may exist.

The critical threshold in the extraction of common property resources is even more interesting. Since household decisions on the extraction of the common

property resource are likely to be affected by expectations of other households' decisions, the strategic response and peer effects may create vicious cycles and multiple equilibria, as explored in the recent exploding literature across many disciplines on social interactions. Further, the interplay of population dynamics with these critical thresholds can change an existing vicious cycle. For example, fertility can modify Dasgupta's model on nutritional status and capacity to work. Poor nutritional status leading to higher infant mortality causes high fertility due to the "insurance of birth." With diminishing returns to labor, high fertility lowers labor productivity, contributing to lower nutritional status and higher mortality rates, *further increasing the motivation for higher fertility*. On the other hand, with abundant land, if production is within the range of increasing returns (e.g., nineteenth-century US Western frontier), high fertility can reverse the initial vicious cycle model of nutritional status and labor productivity. In a capital-scarce environment, greater labor intensification increases fertility through raising the demand for farm labor. High fertility, in turn, creates further demand for food. Further intensification contributes to land degradation, causing a declining resource base and further poverty. De Serbinin et al. (2008) describe the literature on the relationship between fertility and the resource base – farm size, cattle and land, water, etc. Consistent with the vicious cycle hypothesis, they find the relationship to be negative. In contrast, some other authors have found the relationship to be positive. The postulated hypothesis is that households with larger land size have a greater demand for children to retain use rights of the land.

Common property resources provide another example of vicious cycles. In South Asia and sub-Saharan Africa, rural households lack access to tap water or energy sources. They derive a considerable portion of their livelihood resources (firewood, timber, non-timber forest products, fish, bush meat, water, etc.) from common property natural resources. Households make decisions on childbearing as well as labor allocation and marketing. Apart from the intrinsic benefits, the material benefits of children in household goods production are compared against costs of birthing and child maintenance. When a substantial portion of livelihood comes from common property resources, a household's share of the common property depends on the number of hands it can employ to convert common property to private property. Empirical studies from Nepal, Pakistan, and South Africa find evidence supporting the positive relationship between fertility and resource dependency. While having a large number of children exploiting the commons is marginally optimal from an individual household's point of view, it is not optimal for the community. Greater entry into common access resources may lower labor productivity, which households try to offset by adding more hands. This creates incentives for greater household size and higher fertility. But since everyone is doing the same – without consideration to the effect of each household adding more hands on the average productivity and sustainability of resources – degradation accelerates. Environmental resources often have a critical threshold point at which the ecological system loses regenerative capacity, a phenomenon which the ecologists call "loss of resilience". The concept of the resilience has been used in two senses in the environmental science literature. The first one called

“engineering resilience” is defined by the time it takes to return to an equilibrium or steady state after the shock. We here refer to “ecological or Hollings resilience” which presumes the existence of multiple regimes or equilibria. The “loss of resilience” in the latter sense means that the ecological system moves to another equilibrium and no longer retains its former function, structure, identity, and feedbacks. Caught in the fertility-environmental degradation feedback loop, when competition becomes intense to extract resources before others do, the ecological system would gravitate into the catastrophic equilibrium point of “loss of resilience”.

The mediating variables approach is built on recognition that in many regions of the world, the range of choices and trade-offs available to low-income households is affected by the quality and state of the surrounding environment on which their livelihood depends. Thus, informal institutions such as kinship, social norms, and culture and formal institutions (e.g., well-functioning markets for land, labor capital, insurance) are important mediating variables in the population-environment nexus. In sub-Saharan Africa, fosterage within the kinship group has long been viewed as a determinant of high fertility because the costs of responsibility of children diffuse among kinship. On the positive side, Elinor Ostrom shows that traditional communities often protect their local commons from overexploitation by relying on social norms that put restraint on individual community members. Here, social norms may resolve the “coordination problem.”

Formal institutions are often thought to have subverted the traditional informal institutions of common resource management. The emergence of modern governments is often believed to have taken away the authority of villages to impose sanctions against those who violate locally instituted norms of use. As social norms degrade, parents pass off some of the cost of raising children to the community by overexploiting the commons. Baland and Platteau (1996) document the cases of several Sahel states, exerting detrimental impacts on traditional resource management practices. Access to formal markets can affect resource degradation both ways. On the one hand, access can increase resource-depleting productive activities due to higher prices for goods. On the other hand, increases in real wages may increase the harvesting cost of common property.

Relationship Between Migration and the Local Environment. Current research focuses on the relationship between the local environmental condition and household-level decisions to migrate (de Sherbinin et al. 2008), with a focus on growing resource scarcity and ability to access new resources elsewhere. Although more studies are needed to understand the impact of resource scarcity on a household’s decision to migrate, historical examples suggest that scarcity of land resources leads to waves of out-migration to new land. Besides European history, there are more recent instances of core-periphery movement in the developing countries. Examples of core-periphery movement include movement from regions of Brazil and the Ecuadorian Andes to the Amazon. Migration to rural frontiers can give rise to new rural frontiers, utilizing valuable resources in the process; households settled in the first wave of frontier migration use up resources and then send younger members to settle in more distant areas, with the potential for the same pattern repeating itself in the next generation.

Empirical research on the relationships between migration and the environment shows mixed results. Henry et al. (2004) show that in Burkina Faso, the risk of migration is higher in villages with unfavorable agroclimatic conditions. They also find that villages with increased water conservation technologies have a higher likelihood of out-migration. These effects are supposed to be more pronounced for short-term migration as a strategy to diversify income sources in risky environments. Land scarcity has also been shown to be a key driver of migration in Uganda and Nepal. Too unfavorable environmental conditions can also hinder migration; the cost of migration is deemed as an investment, and severe resource constraints may limit the ability to make the investment.

Remittances may have beneficial impact on local environment if the purchased goods are substituted for the goods extracted from local environment. They may also be invested in resource conservation. Farming system is one mediating variable that influences the impact of migration on environmental outcome. Impacts may be less significant in cattle-raising systems where labor demands are small. If there is no functioning credit and insurance market, migration may be adopted as a strategy to mitigate risk. On the other hand, Van Wey (2005) finds that both a lack of land and a large amount of land can motivate migration in Thailand and Mexico. Van Wey (2005) suggests that individuals from households with large landholdings migrate to access capital for investments in technology and other agricultural inputs.

As countries complete the demographic transition with lower fertility and mortality, migration becomes an increasingly dominant force in demographic change. The traditional focus has been at higher levels (aggregate or macro/meso levels) and only on the impact on the destination environment. But more recent work has focused on the “push” created by environmental degradation at the origin and impacts on the origin itself.

Mobility and Population Pressure at the Macro and Meso Levels. The well-cited work of Wilbur Zelinsky builds on the underlying theory of human migration described by Ravenstein, Thomas, Stouffer, and Lee (see Zelinsky 1971 and references therein). Zelinsky (1971) explores the temporal and spatial dimensions of transition in human mobility hypothesized to happen with modernization at macro/meso levels. Zelinsky’s mobility transition model hypothesizes five stages in the transition. Hugo (2011) provides a useful review, criticizing models like Zelinsky’s for failure to consider the two-way migration that is often observed, by focusing on net migration.

Spatial differentials across regions likely contribute to what Hugo (2011) recently described as the “increase in the scale and complexity of both internal and international movement over time” (Hugo 2011, p. S23). He observes that the middle stages of the demographic transition tend to correlate with international migration and with rural-to-urban migration. Further, in these stages – characteristic of many of the world’s developing regions today – young adults tend to be found in higher concentrations in the population and typically have a higher likelihood of migrating compared to older age cohorts (see Hugo 2011 for discussion of other demographic effects).

From an environmental perspective, out-migration to more resource-rich locations with lower population densities may reduce population pressure (and environmental stress) at the origin. However, the regional disruptions that can occur as a result of out-migration may mean disruption of local social norms which keep in check bad environmental behaviors. On the other hand, out-migration accompanied by return migration can bring in new ideas and stimulate innovation, including new environmentally sustainable technologies. The higher degree of temporary mobility that is documented – for example, “international shuttling” behaviors between Mexico and the USA and also the growing density of commuting networks in many regions (Goetz et al. 2010) – may work as a diffusion of innovation generator. Also, the interaction between new (e.g., mobile phone) technologies that generate product demand (to be “like the Jones” in the more-developed world) and the truly massive flow of international migration remittances to the origin may stimulate consumption at the origin. What is clear from a review of the literature is that population growth, environment, and migration are likely interrelated but deserve much more attention.

Rural-Urban Adjustments. Finally, no review of the population-environment nexus would be complete without mention of the growing prominence of urban places across the globe. Urban places have become particularly attractive population magnets; the global urban population has now surpassed the global population defined as rural. This reflects domestic and international migration fueled by GDP concentrations in city centers. The trend appears to be worldwide; even in South and Southeast Asia, Africa, and Latin America, humans are rapidly concentrating into urban centers (PRB 2010). Long-term trends show that the US population but also per capita GDP have become even more concentrated in metro areas, an international trend shared by the OECD countries and most countries worldwide.

Rapid urbanization “hinders the development of adequate infrastructure and regulatory mechanisms for coping with pollution and other byproducts of growth, often resulting in high levels of air and water pollution and other environmental ills” (Hunter 2000, p. xiii). Urbanization can also alter “local climate patterns” (Hunter 2000), with concentrations of artificial surfaces creating heat islands (Hunter 2000). Another common result is sprawl. Unless checked by geography or policy, population eventually spreads over the surrounding landscape. In some cases, consumption drives the transition. Greater mobility of the human population also contributes to the spillover onto nearby landscapes, for consumption of natural resource amenities, for perceived healthier lifestyles and safety, but also for lower costs of living. Environmental amenities in the countryside are recognized as strong attractants (see Cherry and Rickman 2011).

Population growth and greater mobility, higher consumption levels in some places, and the age distribution of the population are expected to influence patterns of rural-exurban-urban transformation and rates of transition of the landscape and ecosystem services. Important related issues include the following: food production for growing urban populations, provision of specific ecosystem services, and use of natural resources including land proximate to population centers. Food availability and regional and local food production capacity represent major concerns, with access to food and water in the less- and least-developed regions being paramount.

Finally, while countries are at different stages in this process, the process itself appears to be essentially the same (Findeis et al. 2009). That is, the population-driven changes taking place across exurban or peri-urban landscapes are in some respects very similar worldwide.

55.5 Conclusions

The total population living on earth continues to grow rapidly, putting significant pressure on the earth's ecosystem. While the rate of growth has recently declined, absolute numbers of people sharing the earth's space continue to grow. Technological change (e.g., sustainable agricultural technologies), input substitution (e.g., substitution of biopower for fossil fuels), and the emergence of new institutions will contribute to reduce the environment problem in the developed countries, although impacts on the environment stemming from affluence and high levels of consumption will remain issues. Further, transboundary externalities – the shifting of the costs of economic development from the “haves” to the “have nots” but also among the “haves” – will continue to take center stage in environmental discourse. Well-known issues include climate change, loss of biodiversity, and water quality and quantity, among other well-publicized environmental issues.

The less- and least-developed regions of the world will face particularly stiff challenges. Population growth rates are projected to be highest in these regions. As argued in this chapter, some of what are considered the most vulnerable places on earth are inhabited by populations with high population growth rates and lower chances of developing economically for reasons largely stemming from their environment and geography. The major challenge will be to be resilient in the face of the dual challenges of own growth and the force of transboundary externalities created by others. Innovation should help to reduce impacts, but whether innovation will be targeted to sustainable solutions adaptable to these regions is a critical question.

For regional science, the challenge will be greater focus on three issues: (i) the interface between the earth's more-developed and less-/least-developed regions, and how to reduce environmental impacts stemming from the developed world; (ii) development within the less-/least-developed regions of the globe to reduce environmental impact as growth occurs; and (iii) the three-way interaction between population, environment, and migration. Population change was identified recently by the Social, Behavioral and Economic Sciences Directorate of the National Science Foundation as one of four major topic areas for future research (NSF 2011). “Sources of disparities” is the second of the four areas identified for major emphasis. Regional scientists can provide insight into both major challenges but especially deepen our understanding of how to effectively reduce differentials between affluent and poor regions of the world in this century, a challenge of paramount importance.

As shown in this chapter, the incentives, thresholds, and nonlinear relationships surrounding population change and sources of disparities are highly complex.

In many respects, we are only in the early stages of understanding the complex underlying mechanisms and relationships behind a full understanding of the population-environment nexus. Evolution of the interrelationships among regions, a purview of regional science, is known for its complexity. Also complex is how to balance – *over a very short period of time* – population growth and food, water, and other resource requirements by humans in different regions of the world. Hunter (2000) argues that we need a more “precise scientific understanding of the complex interactions between demographic processes and the environment” (Hunter 2000, p. xx in report’s foreword). Scientists, including regional scientists, will need to collect robust data and develop new models that link natural and social-economic-behavioral processes and build a compelling library of credible evidence from across the globe to inform decision-making and governance at multiple scales. A major challenge will be coordinating this work across the multiple scales and with multiple disciplines and publics to avoid what could be catastrophic impacts. However, there is really no choice.

References

- Baland JM, Platteau JP (1996) Halting degradation of natural resources: is there a role for rural communities? Clarendon, Oxford
- Birdsall N (1988) Economic approaches to population growth. In: Chenery H, Srinivasan TN (eds) Handbook of development economics, vol 1. Elsevier, Amsterdam, pp 477–542, Chapter 12
- Campbell M (2007) Why the silence on population? *Popul Environ* 28(4):237–246
- Carson RT (2010) The environmental Kuznets curve: seeking empirical regularity and theoretical structure. *Rev Environ Econ Policy* 4(1):3–23
- Cherry T, Rickman D (2011) Environmental amenities and regional economic development. Routledge, London
- Coale AJ, Hoover EM (1958) Population growth and economic development in low-income countries: a case study of India’s prospects. Princeton, NJ: Princeton University Press
- Dasgupta P (2003) Population, poverty, and the natural environment. In: Maler KG, Vincent JR (eds) Handbook of environmental economics, vol 1, 1st edn. Elsevier, Amsterdam, pp 191–247, Chapter 5
- de Sherbinin A, VanWey LK et al (2008) Rural household demographics, livelihoods and the environment. *Glob Environ Chang* 18(1):38–53
- Dietz T, Rosa EA, York R (2007) Driving the human ecological footprint. *Front Ecol Environ* 5(1):13–18
- Edwards R (2011) Changes in world inequality in length of life: 1970–2000. *Popul Dev Rev* 37(3):499–528
- Ehrlich PR, Holdren JP (1971) Impact of population growth. *Science* 171(3977):1212–1217
- Findeis J, Brasier K, Salcedo Du Bois R (2009) Demographic change and land use transitions. In: Goetz S, Brouwer F (eds) New perspectives on agri-environmental policies: a multidisciplinary and transatlantic approach, Chapter 2. Routledge, Abingdon, pp 13–40, Chapter 2
- Gallup JL, Sachs JD, Mellinger AD (1999) Geography and economic development. *Int Reg Sci Rev* 22(2):179–232
- Galor O (2005) From stagnation to growth: unified growth theory. In: Aghion P, Durlauf S (eds) Handbook of economic growth, vol 1, 1st edn, Chapter 4. Elsevier, Amsterdam, pp 171–293
- Goetz S, Han Y, Findeis J, Brasier K (2010) US commuting networks and economic growth: measurement and implications for spatial policy. *Growth Change* 41(2):276–302

- Grossman GM, Krueger AB (1995) Economic growth and the environment. *Q J Econ* 110(2):353–377
- Henry S, Schoumaker B et al (2004) The impact of rainfall on the first out-migration: a multi-level event-history analysis in Burkina Faso. *Popul Environ* 25(5):423–460
- Hugo G (2011) Future demographic change and its interactions with migration and climate change. *Glob Environ Chang* 215(Supp 1):S21–S33
- Hunter LM (2000) The environmental implications of population dynamics. RAND, Santa Monica, 33 pp
- McNeill JR (2006) Population and the natural environment: trends challenges. *Popul Dev Rev* 32(suppl):183–202
- NIC (National Intelligence Council) (2008) Global trends 2025: a transformed world. NIC 2008-003, Washington, DC
- Nordhaus WD (1973) World Dynamics: Measurement Without Data. *Econ J* 83(December): 1156–1183.
- NSF (National Science Foundation) (2011) Rebuilding the mosaic: fostering research in the social, behavioral, and economic sciences at the National Science Foundation in the next decade. NSF 11-086, Arlington 63 pp
- Popp D (2002) Induced innovation and energy prices. *Am Econ Rev* 92(1):160–180
- PRB (Population Reference Bureau) (2010) 2010 World population data sheet. http://www.prb.org/pdf10/10wpds_eng.pdf. Accessed 1 Jun 2012
- Scherbov S, Lutz W, Sanderson WC (2011) The uncertain timing of reaching 8 billion, peak world population, and other demographic milestones. *Popul Dev Rev* 37(3):571–578
- UNFPA (United Nations Population Fund) (2011) The state of world population 2011: people and possibilities in a world of 7 billion. United Nations Population Fund, New York, 123 pp
- United Nations, Department of Economic and Social Affairs, Population Division (2011) World population prospects, the 2010 revision. United Nations, New York
- Van Wey LK (2005) Land ownership as a determinant of international and internal migration in Mexico and internal migration in Thailand. *Int Migr Rev* 39(1):141–172
- Vitousek PM, Mooney HA, Lubchenco J, Melillo JM (1997) Human domination of Earth's ecosystems. *Science* 277(5325):494–499
- Zelinsky W (1971) The hypothesis of the mobility transition. *Geogr Rev* 61(2):219–249

Section VII

Spatial Analysis and Geocomputation

spatial space-time
information geographic applications
analysis function
MAUP correlation
scale
social simulation applied
probability
networks values
population constraint
constraint
values
time locations
visual
constraint
location
GIS urban
distance
Variables
model Openshaw
area
data estimates
geocomputation
local
mean
change
methods distribution
map techniques local
area mean
microsimulation change
behavior GIScience
spatio-temporal systems network aggregation

Michael F. Goodchild and Paul A. Longley

Contents

56.1	Introduction	1108
56.2	Principles of GIScience	1110
56.2.1	The Characteristics of Geographic Information	1110
56.2.2	Dealing with Large Data Volumes	1112
56.2.3	Scale-Related Issues	1113
56.2.4	Simulation in GIScience	1114
56.2.5	Achievements of GIScience	1115
56.3	Changing Practice and Changing Problems	1116
56.3.1	CyberGIS and Parallel Processing	1117
56.3.2	The Social Context of GIS	1118
56.3.3	Neogeography, Wikification, and Open Data	1119
56.4	Conclusion	1121
	References	1121

Abstract

This chapter begins with definitions of geographic information science (GIScience), of geocomputation, and of spatial analysis. We then discuss how these research areas have been influenced by recent developments in computing and data-intensive analysis, before setting out their core organizing principles from a practical perspective. The following section reflects on the key characteristics of geographic information, the problems posed by large data volumes, the relevance of

M.F. Goodchild (✉)

Center for Spatial Studies and Department of Geography, University of California, Santa Barbara,
CA, USA

e-mail: good@geog.ucsb.edu

P.A. Longley

Department of Geography, University College London, London, UK

e-mail: p.longley@geog.ucl.ac.uk

geographic scale, the remit of geographic simulation, and the key achievements of GIScience and geocomputation to date. Our subsequent review of changing scientific practices and the changing problems facing scientists addresses developments in high-performance computing, heightened awareness of the social context of GIS, and the importance of neogeography in providing new data sources and in driving the need for new techniques.

56.1 Introduction

Geographic information science (GIScience) addresses fundamental issues associated with geographic information and the use of geographic information systems to perform spatial analysis, using a scientific approach (for detailed discussions of the nature of geographic information science, see Duckham et al. 2003). The issues may be practical, as in the question of how to address uncertainty in geographic information; they may be empirical, as in the observation generally known as Tobler's First Law of Geography ("All things are related, but nearby things are more related than distant things.") (Tobler 1970); or they may be theoretical, as in the fundamental contribution known as the 9-intersection of topology (Briefly, the set of topologically distinct relationships that can exist between two areas in the plane) (Egenhofer and Franzosa 1991). To some, the term implies the use of geographic information systems (GIS) as a scientific tool in research and decision-making, and as such it has been widely applied to the solution of virtually any problem that is embedded in geographic space, from global warming to crime and water pollution. Much progress has been made in GIScience in the two decades since the term was coined (Goodchild 1992), through the efforts of a growing scientific community. It is also important to note that other terms convey similar meaning, including geomatics, geoinformatics, and spatial information science, and that GIScience plays an important role in the practice of regional science, both as a technology that can support research and as an approach to problem-solving.

Geocomputation is also fundamentally concerned with geographic information, in other words information about features and phenomena and their locations on or near the Earth's surface. Coined a little later by Openshaw and Abrahart (1996), the term is often used in cross-sectional analysis to describe the repeated analysis and simulation of spatial distributions, in order to explore spatial distributions and to draw inferences about them. More specifically, the term is often taken to imply simulation of processes operating in the geographic domain and thus with geographic information that is primarily dynamic. The major issues in geocomputation often center on the computational problems that arise in simulating complex systems with massive numbers of features, data items, or agents. In this sense geocomputation develops an application-led focus upon the way the world *works*, founded upon rich digital representations of the way that the world *looks*, and makes prediction a central goal. The main contribution of geocomputation may thus lie in the development of better tools for dealing with complex, dynamic systems.

From these definitions it is clear that GIScience and geocomputation have much in common, that their interests overlap substantially, and that it may even be helpful to think of geocomputation as a computationally intensive, application-led component of GIScience. Accordingly, the focus of this chapter is on the common ground between them, using the terms somewhat interchangeably. The term GIScience is used wherever the context seems to demand it and similarly with the term geocomputation. Both terms are fundamentally concerned with spatial analysis, defined as the set of methods whose results change in response to changes in the locations of the objects being analyzed, and we sometimes use this umbrella term. The remainder of this section elaborates on the basic definition of GIScience and the research conducted under its banner. This is followed by a discussion of the basic principles of GIScience; in a nod to geocomputation, the discussion emphasizes those areas where GIScience has been successful at solving computationally intensive problems. Major methods of analysis are reviewed. The third section of the chapter addresses changing practices in GIScience, focusing on the increasing importance of collaboration, on novel data sources, and on the problems of dealing with uncertainty. Science generally is changing in response to the need to study complex systems and the use of simulation, and this trend is certainly affecting GIScience. The concept of data-intensive science, the so-called *Fourth Paradigm* (Hey et al. 2009), has a natural fit to geographic problems and their massive volumes of data, while the meta-issues of documentation and provenance are beginning to loom large in a science that is no longer dominated by the individual investigator. Finally, the fourth major section speculates on the future and discusses the coevolution of GIScience and geocomputation. Future developments are likely to be driven, as in the past, by trends in data, in computation, and in the society that forms the context for both fields.

While debates about the nature and meaning of *science* have raged for centuries and will probably never end, the core ideas are clear. First, science seeks laws and principles that can be shown to be valid in the observable world and are generalizable in the sense that they apply everywhere and at all times. Both of the examples cited earlier – Tobler’s First Law and the 9-intersection – are clearly of this nature, and as a theoretical conclusion, the 9-intersection not only applies everywhere at all times but also applies in any imaginable space. Second, science is founded on definitions of terms that are rigorously stated and understood by all scientists. Third, scientific experiments and their results are replicable, being stated in sufficient detail that someone else could expect to obtain them by carrying out an identical experiment. In this context the term *black box* is pejorative since procedures that are hidden inside a box cannot be described and therefore cannot be replicated. Well-understood principles also apply to the details of reporting, as in the rule that any measurement or numerical result be stated to a precision (number of significant digits) that reflects the accuracy of the measuring device or model. Principles such as these help define GIScience and geocomputation and distinguish them from less rigorous applications of GIS and related technologies.

A distinction is often drawn between *pure* science, or science for the sake of curiosity and the quest for general discoveries, and *applied* science, or science that aims to solve problems in the observable world using scientific methods.

The geo- prefix reminds us that the Earth provides a unique laboratory for scientific investigation, and the uniqueness of the places on it often limits the scope for the kinds of controlled experiments that characterize scientific activity in other disciplines. Geographic space is the space of human activity, and most of the problems human society is facing are embedded in it, from poverty and hunger to health. Indeed, it is hard sometimes to avoid application in GIScience because the field is inevitably close to the real world, a fact that perhaps accounts for at least some of the passion displayed by its practitioners. Moreover, curiosity has often provided the motivation to explore, characterize, and map the geographic world, though the results of such exploration are rarely generalizable in the sense that Newton's laws of motion or the Mendeleev periodic table are generalizable.

This pure/applied distinction explains how progress in spatial analysis is measured. On the one hand, the refereed journals in which much successful GIScience research is published and the presentations at conferences such as the biennial International Symposia on Geographic Information Science emphasize the purer forms of science, while other conferences, such as the biennial International Conferences on Geocomputation, emphasize how the core organizing principles and concepts of GIScience can be brought to bear on solving practical problems. A large industry, valued according to some estimates as \$20 billion annually (Longley et al. 2011), has sprung up around the data acquisitions and tools needed in such practical problem-solving. Clearly the metrics of success here are much more diverse than in pure science.

56.2 Principles of GIScience

In this section we describe some of the major achievements of GIScience in its first two decades. The selection includes advances that closely resemble geocomputation in the sense of being concerned with large, complex systems and with large volumes of data. We begin with a discussion of the characteristics that distinguish geographic information and geographic problem-solving from data-driven science in other domains. We then discuss the strategies that have been adopted in GIScience for avoiding or successfully dealing with the problems of large data volumes, including aggregation, divide and conquer, and compression. We discuss some of the unintended consequences of such strategies, in the form of uncertainty, the ecological fallacy, and the modifiable areal unit problem. We elaborate on the nature of simulation in geographic space, on some of the more successful research conducted in this area, and on some of the issues it raises. Finally, we present a brief summary of progress in GIScience in the past 20 years.

56.2.1 The Characteristics of Geographic Information

One of the first attempts to identify the special characteristics of geographic information, or *what is special about spatial?*, was made by Anselin (1989).

He argued that two characteristics were universal: spatial dependence and spatial heterogeneity. Reference has already been made to the first, in the form of Tobler's First Law of Geography: "All things are similar, but nearby things are more similar than distant things." While we can argue about whether the statement meets the criteria for a law as that term is normally understood by philosophers of science and whether exceptions should be allowed, it is clear that the vast majority of phenomena distributed over the Earth's surface and near surface adheres to it while differing in precisely how similarity decays with distance. Moreover, there is no doubt of the law's efficacy in GIS.

The principle is essentially one of context, since it requires a phenomenon at one point to be consistent with the same phenomenon at nearby points. It appears to apply well in three-dimensional space and also to apply in four-dimensional space-time. Perhaps the easiest way to demonstrate its validity is by a thought experiment in which it is not true, where a minute displacement on the Earth's surface produces a completely independent environment – clearly this does not happen and cannot happen except in rare circumstances.

As a cornerstone of GIScience, the principle has two major implications. First, similarity over short distances allows the Earth's surface to be divided into regions within which phenomena are approximately homogeneous, achieving great economies in data volume by expressing attributes as properties of entire areas rather than of individual points. In short, the principle enables the assumed-homogeneous polygons that dominate many representations in GIS. Similarly, it allows reasonable guesses to be made of the properties of places that have not been visited or measured, in a process known as spatial interpolation. The principle thus justifies the techniques that are used, for example, to create weather maps from scattered point observations.

Unfortunately the principle of spatial dependence also provides a major headache for researchers working with geographic information, since it runs counter to the assumption made in many statistical tests that the data were acquired through a process of random and independent sampling from a parent population. An analysis of the 58 counties of California, for example, cannot make that assumption since the principle implies that conditions in neighboring counties will be similar. Moreover, there is no larger universe of which the set of all counties of California constitute a random sample.

Anselin's second principle addresses spatial heterogeneity, or the tendency for parts of the Earth's surface to be distinct from one another. This also has profound implications. Consider, for example, a local agency seeking to define a taxonomy of local land use. The result will inevitably be different depending on the agency's location and the local conditions in its jurisdiction, and every jurisdiction will argue that its scheme is better than any global or national standard. In early geodesy, the figure of the Earth (the mathematical function used to approximate the Earth's shape and thus define latitude and longitude) was unique to each jurisdiction or region, and it was not until the 1960s that pressure for a single standard prevailed, driven by the growing importance of air travel and the targeting of intercontinental ballistic missiles. Unfortunately any universal standard will inevitably be

suboptimal for any local jurisdiction, whether it be over land-use classification or the shape of the Earth, so there will always be tension between the desire to be locally optimal and the desire to be globally universal.

56.2.2 Dealing with Large Data Volumes

The previous section was concerned with principles that can be demonstrated to be empirically true. We now move to a discussion of some of the principles that guide the design of GIS technology and allow GIS to deal with problems that might otherwise be overwhelmingly voluminous, a key issue in geocomputation given its goal of addressing large problems. The Earth's surface has approximately 500 million square kilometers, and a description of it at a resolution of 1 m² would therefore create 500 trillion data elements if no strategy were adopted to reduce the volume. Even allocating a single byte to each data element would create half a petabyte of data.

In the previous section we discussed Tobler's First Law, the basis for aggregating data elements into statements about entire polygons. California's land area amounts to 403,800 km², and describing each sq m with a two-byte designation CA would produce roughly 0.8 terabytes of data. But capturing the coordinates of its boundary and adding a single attribute CA to the polygon could clearly compress this to only a few kilobytes, even with precise coordinates, and by recording only a single attribute would avoid the potential for error in the vast number of identical attributes that would have to be recorded in the raster approach. Alternatively, a variety of compression techniques can also be used to replace a raster of individual data elements with a series of <run-length, value> pairs. Many other methods of compression, generalization, and abstraction have been devised to deal with the volume problem, some of them *lossy* in the sense that the result is only approximately identical and the original data cannot be recovered from the compressed version and some of them *loss-less*.

In a *divide-and-conquer* strategy, a geographic area is partitioned, and analysis or modeling proceeds one partition at a time. The term *tile* is often used for partition, especially where the partitions are rectangular. Instead of solving a problem for the whole of California, for example, one might solve it separately for each of its counties. Interactions exist between counties in almost every application: in analyzing water pollution, for example, the actions of a county will influence the water quality in any downstream county, and air pollution will travel to any counties downwind. Thus, a successful divide-and-conquer strategy must also consider the degree to which counties interact and include this in the model, often by iterating between modeling within-county effects and modeling between-county effects. Nevertheless, the overall computational efficiency of the modeling will probably be improved by adopting this strategy. Many GIS algorithms make explicit use of divide and conquer, as an approach to handling the vast amounts of data provided by satellite-based remote sensing, and implicit divide and conquer has been an intrinsic part of human problem-solving from time immemorial.

56.2.3 Scale-Related Issues

The term *scale* is often used in GIScience in the sense of spatial resolution, to distinguish between fine-scale or detailed data and coarse-scale or generalized data. Some of the techniques described in the previous section essentially sacrifice scale in the interests of reducing data volume. To a cartographer, reducing a map's *representative fraction*, the ratio of distance on the map to distance on the Earth, is similarly a sacrifice of scale, often in the interests of visual clarity. To a compiler of social statistics, reporting counts of people based on large, aggregated reporting zones may also be a means of reducing data volume.

All of these techniques have consequences that are well recognized in GIScience. The *modifiable areal unit problem* refers to the effects that changes in reporting zone boundaries will have on the results of any geographic analysis. The term was first formally characterized by Openshaw (1983), who demonstrated that changing reporting-zone boundaries could produce dramatic swings in results, even when holding scale constant. His solution, which became a fundamental tenet of geocomputation, was to recommend exploring the aggregation effect in any specific case, by repeated analysis using different zones. Unfortunately in most cases, this can only be done by aggregating predefined zones, producing different results but at a still coarser level of aggregation, since data compiled for different zonal arrangements at the same level of aggregation will usually not be available. Many studies have documented the problem, while others have argued that it results not from a failure of analytic method but from a failure on the part of the investigator to be explicit about the scale at which the hypothesized effects occur. For example, in Openshaw's original example, the 99 counties of Iowa were used to explore the relationship between percent of the population over 65 and percent registered Republican voters. Aggregating the counties in various ways did indeed produce different results but at coarser scale. What is missing in this case is a well-defined hypothesis as to why this correlation should appear and at what scale. Perhaps the process works at the individual level, and older people are more likely to vote Republican, in which case the hypothesis is best tested at the individual level. Or perhaps the process is ecological: a neighborhood with a large percent of people over 65 also attracts a large percent of Republican voters, whether or not they are over 65. In the latter case, the appropriate scale of analysis is that of the neighborhood, requiring a formal definition of that concept and an aggregation of fine-scale data, such as block-group data, to the neighborhood level. The general point is relevant to the definition of spatial analysis in Sect. 56.1 — that we should not be looking for statistics that are invariant to the phenomenon that we wish to study. As such, the MAUP is not an empirical problem but rather is a theoretical requirement to hone statistics to the geographic context in which they are applied.

A closely related problem, also well recognized in GIScience, is the ecological fallacy, the fallacy of reasoning from the aggregate to the individual. The fallacy already appeared in the previous paragraph, since it would be wrong to infer from a county-level correlation that individuals over 65 tend to vote Republican – in fact, in the extreme, Openshaw's correlations could exist in Iowa at the county level even

though no person over 65 was a registered Republican. King (1997) reviews the problem in greater detail and suggests ways of addressing it. Other approaches to downscaling, or replacement of coarse-scale data by fine-scale data, can be found, such as the work of Boucher and Kyriakidis (2006) in the context of remote sensing.

56.2.4 Simulation in GIScience

Many processes that operate on the Earth's surface can be abstracted in the form of simple rules. One might hypothesize, for example, that consumers always purchase groceries from the store that can be reached in minimum time from their homes. Exactly how such hypotheses play out in the real world can be difficult to predict, however, because of the basic heterogeneity and complexity of the Earth's surface. Christaller was able to show that such simple assumptions about behavior led to simple patterns of settlements in areas dominated by agriculture, but only by assuming a perfectly uniform plane. Similarly, Davis was able to theorize about the development of topography through the process of erosion, but only by assuming a starting condition of a flat, uplifted block. Research in both areas has clearly demonstrated that the perfect theoretical patterns predicted never arise in practice.

One strategy for addressing such issues is to assume that in the infinite complexity of the real world, all patterns are equally likely to emerge, and that the properties we will observe will be those that are most likely. This strategy enabled Wilson (1970) to show that the most likely form of distance decay in human interaction was the negative exponential, and Shreve (1966) was able to show that the effect of random development of stream networks would be the laws previously observed by Horton. Similar approaches have been applied to the statistical distribution of city size or the patterning of urban form (Batty and Longley 1994).

Nevertheless, while they yield results that are often strikingly in agreement with reality, such approaches lack the practical value that real-world decision-making demands. Instead, GIScience and geocomputation are increasingly being used to simulate the effects of simple hypotheses about behavior on the complex landscapes presented by the geographic world. The generality of such approaches lies in the hypotheses they make about behavior; the landscapes they address, and the patterns they produce, are essentially unique.

Such approaches fall into two major categories, depending on how the hypotheses about behavior are expressed. The approach of *cellular automata* begins with a representation of the landscape as a raster and implements a set of rules about the conditions in any cell of the raster. The approach was originally popularized by Conway in his Game of Life, in which he was able to show that distinct patterns emerged through the playing out of simple rules on a uniform landscape. Such patterns are known as *emergent properties*, since they would be virtually impossible to predict through mathematical analysis. The cellular-automata approach has been used by Clarke (e.g., Clarke and Gaydos 1998) and others to simulate urban growth, based on simple rules that govern whether or not a cell will

change state from undeveloped to developed. Such approaches allow for the testing of policy options, expressed in the form of modifications to the rules or to the landscape, and have been widely adopted by urban planners.

The alternative approach centers on the concept of *agent*, an entity that is able to move across the geographic landscape and behave according to specified rules. This *agent-based* approach is thus somewhat distinct from the cell-based approach of cellular automata. Agent-based models have been widely implemented in GIScience and geocomputation. For example, Torrens, Li, and Griffin (2011) have studied the behavior of crowds using simple rules of individual behavior, with applications in the management of large crowds with their potential for panic and mass injury. Evans and Kelley (2004) have studied the behavior of decision-makers in their role in the evolution of rural landscapes and examined policies that may lead to less fragmentation of land cover and thus greater sustainability of wildlife. Maguire, Batty, and Goodchild (2005) discuss several other examples of cellular automata and agent-based models in GIScience and geocomputation.

Both approaches raise a number of issues (for a general discussion of these issues, see, e.g., Parker et al. 2003). From an epistemological perspective, several authors have explored the role of such modeling efforts in advancing scientific knowledge. On the one hand, a model is only as good as the rules and hypotheses about behavior on which it is based. It is unlikely that the results of simulation will lead directly to a modification of the rules and more likely that rules will be improved through controlled experiments outside the context of the modeling. If patterns emerge that were unexpected, one might argue that scientific knowledge has advanced, but on the other hand, such patterns may be due to the specific details of the modeling and may not replicate anything that actually happens in the real world.

Validation and verification of simulation models are always problematic, since the results purport to represent a future that is still to come. *Hindcasting* is a useful technique, in which the model is used to predict what is already part of the historic record, usually by working forward from some time in the past. But the predictions of the model will never replicate reality perfectly, forcing the investigator to ask what level of error in prediction is acceptable and what is unacceptable. Moreover, it is possible and indeed likely that rules and hypotheses about social behavior that drive the model will change in the future. In that regard, models of physical processes may be more reliable than models of social processes.

56.2.5 Achievements of GIScience

As we noted earlier, the term *GIScience* was coined in a 1992 paper (Goodchild 1992). In some ways the paper was a reaction to comments being made in the literature about the significance of GIS that it was little more than a tool and did not therefore deserve a place in the academy. The funding of the US National Center for Geographic Information and Analysis (NCGIA) in 1988 by the National Science Foundation seemed to indicate a willingness in some quarters to see more in GIS

than technique. Nevertheless the tool/science debate continued for some time and is summarized by Wright, Goodchild, and Proctor (1997).

Two decades later, several efforts were made to look back and assess progress. A meeting for that purpose was convened in Santa Barbara in December 2008 (<http://ncgia.ucsb.edu/projects/isgis/>), and a paper summarizing its results and offering a personal perspective has been published by Goodchild (2010). It draws on the assessments of several individuals and on a bibliographic analysis performed by Skupin. While any level of consensus is inevitably difficult to achieve, the following might be argued to be the major achievements of two decades of GIScience:

- Clarification and specification of the basic data model, including recognition of the fundamental significance of discrete-object and continuous-field conceptualizations, the emergence of object-oriented data modeling, and the specification of spatial relations
- The development of place-based techniques of spatial analysis, including local indicators of spatial association (Anselin 1995; Ord and Getis 1995), spatial regression models (LeSage and Pace 2009), and geographically weighted regression (Fotheringham et al. 2002)
- The specification of standards for simple features, metadata, real-time interaction across the Internet, and many other aspects of GIS practice, led by the Open Geospatial Consortium and the US Federal Geographic Data Committee
- The development of digital globes such as Google Earth that allow real-time interaction with three-dimensional models of the Earth
- Recognition of the importance of ontology, as the key to interoperability across communities, languages, and cultures
- Search and retrieval based on geographic location, through mechanisms such as the geoportal (Maguire and Longley 2005)
- Advances in geovisualization, going far beyond the capabilities of conventional cartography to include animation, the third spatial dimension, reduction of high-dimensional data sets, and many other topics
- Achievement of a new level of understanding of uncertainty in geographic information, its handling, and its effects, together with a fundamental shift of focus from accuracy to uncertainty

Perhaps more important are the institutional achievements, which can be seen as the indirect result of such advances. GIScience is now widely recognized in the titles of journals and the names of departments and programs. In recent years several GIScientists have been elected to prestigious institutions such as the US National Academy of Sciences and the UK's Royal Society. GIScience conferences have proliferated, and the GIScience bookshelf now contains an impressive array of titles.

56.3 Changing Practice and Changing Problems

In this section we examine the changing nature of GIScience and speculate on its future. GIS has always been driven by competing factors. On the one hand, it has been at the mercy of trends and changes within the larger computing industry,

including new technologies that may or may not offer significant benefits for GIS. For example, the relational database management systems of the 1970s led to a major breakthrough in data modeling in GIS. GIS has also been driven by the need to solve problems of importance to society, from the resource management that provided the initial applications of GIS in the 1980s to the military applications that have always been important but half hidden, and new applications in public health that are as yet only partially developed. GIS as a tool for science is subject to the winds of change that are currently blowing through the scientific community, pushing it toward a more collaborative, multidisciplinary paradigm. Finally, GIS exists in a social context of concerns about privacy and about the role that an expensive technology can play in empowering the already empowered, and is beginning to recognize the importance of the average citizen as both a consumer and producer of geographic information.

This section is structured as follows. We begin with a discussion of high-performance computing and its importance for the kinds of massive simulation models discussed previously. We then move to a discussion of the social context of GIS and the social critique that emerged in the 1990s and now drives the research of many GIScientists. Finally, we examine the phenomenon of *neogeography* and the importance it may hold in providing new data sources and in driving the need for new techniques.

56.3.1 CyberGIS and Parallel Processing

A major report of the US National Science Foundation (NSF 2003) proposed the term *cyberinfrastructure* to describe the kinds of computing infrastructure that would be needed to support science in the future. Instead of the lone investigator and the desktop system, the report envisioned a distributed infrastructure that would support widespread collaboration across a range of disciplines, following the notion that science in the future would address complex problems with complementary teams of scientists of varied expertise. The solution of complex, large-scale problems would also require a heavy level of investment in high-performance computing (HPC) with its massively parallel architectures. Parallel architectures have an inherently good fit to the nature of geographic space and its somewhat independent individual and community agents, all of which can be seen as semi-independent decision-makers acting in parallel rather than serially.

A number of authors have argued that geographic research and problem-solving require a specific form of cyberinfrastructure that addresses several key issues and have coined the term *cyberGIS*. How exactly should the geographic world be partitioned across processors? How should one measure computational intensity as a geographic variable? How should the user interface of an integrated cyberGIS be designed? What types of problems, models, and analyses best justify these new approaches? What incentives will persuade the average GIScientist to engage with cyberGIS, given the initial impression of complexity and inaccessibility and a high level of personal investment in conventional GIS?

Efforts to parallelize GIS date from the 1990s but were not successful for several reasons. First, parallel computing was expensive at the time, and it was difficult for investigators to justify the cost. Second, parallel computing was rendered inaccessible by the need to reprogram in specialized languages. Third, while it was easy to find examples of geographic problems that involved massive volumes of data, it was harder to find ones that involved massive computation. Finally, collaborative technologies had not yet advanced to the point where it was possible for widely distributed research teams to work together productively.

Many of these arguments are now moot, however. HPC is widely available, and Cloud and Grid technologies are making the transition from conventional computing almost transparent. The need for collaboration is much stronger, and the kinds of problems that used to be solved by individual investigators are now hard to find. Finally, geocomputation has opened the doors to the kinds of massive computation that HPC is designed to address. Indeed, the most compelling examples of the need for HPC lie in the kinds of agent-based and cellular simulations reviewed in the previous section.

In recent years it has also become possible to parallelize processing on the desktop, following the addition of graphical processing units (GPUs) to graphics boards in order to improve the quality and speed of image rendering. Although an innovation of the computer games market, GPU chips were subsequently adapted to more general-purpose computing: today, Nvidia (which, along with AMD, is the world's largest graphics-card manufacturer) produces chips designed specifically for non-graphics applications and provides a specialized programming-language architecture for use with them. GPUs outperform traditional computation on a central processing unit (CPU) because a GPU has a higher density of cores and uses a process called streaming to handle a number of operations simultaneously. The result is increased processing speed of computationally intensive algorithms. General-purpose computing on graphics processing units (GPGPU) describes the exploitation of the resources of the GPU for various tasks which might previously have been conducted on a CPU. It has particular advantages for real-time systems where the speed of return of results is fundamental to usability and interaction.

Adnan, Longley, and Singleton ([in press](#)) describe an application in geocomputational geodemographics, in which *k*-means (a frequently used algorithm in the creation of geodemographic classifications) is enhanced to run in parallel over a GPU. This work exploits the parallel-computing computer unified device architecture (CUDA), which allows code written in standard C or C++ to be used in GPU processing.

56.3.2 The Social Context of GIS

Although the GIS technology that underpins GIScience and geocomputation is an established part of the IT mainstream, there is enduring unease in some academic quarters about the social implications of this technology. Early statements were contained in Pickles' ([1995](#)) edited volume *Ground Truth: The Social Implications of Geographic Information Systems*, which remains an enduring statement of

concerns built around four principal issues. First, there is the view that GIS technology is used to portray homogeneity rather than representing the needs and views of minorities and that this arises in part because systems are created and maintained by vested interests in society. The roots to this critique can be traced to a wider debate as to whether the umbrella term GIS is best conceived as a tool or as a science and is something that can be addressed through clarifying the ontologies and epistemologies of GIScience and geocomputation. Second, there is the view that use of a technological tool such as GIS can never be inherently neutral and that GIS is used for ethically questionable purposes, such as surveillance and the gathering of military and industrial intelligence. Web 2.0, discussed below, has begun to address this criticism, since it has gone some way to level the playing field in terms of data access and enabled participation of a wider cross section of society in the use of this technology of problem-solving. Moreover, it is difficult to construe the views of the Earth promulgated through services such as Google and Bing as intrinsically privileged, not least if they are open to all with access to an Internet browser. Third, there has been a dearth of applications of GIS in *critical* research and a preoccupation with the quest for analytical solutions rather than establishing the impacts of human agency and social structures upon unique places. The rise of mixed-method approaches to GIS (Cope and Elwood 2009) has gone some way toward addressing these concerns. Finally, there is still a view in some quarters that GI systems and science are inextricably bound to the philosophy and assumptions of the approach to science known as *logical positivism*. This implies that GIScience in particular, and science in general, can never be more than a positivist tool and a normative instrument and cannot enrich other more critical perspectives in geography. Although still featured in many introductory courses on social science methodologies, this critique is something of a caricature of the positivist methods that pervade scientific investigation more generally.

56.3.3 Neogeography, Wikification, and Open Data

Recent years have seen the reuse of the term *neogeography* to describe the developments in Web mapping technology and spatial data infrastructures that have greatly enhanced our abilities to assemble, share, and interact with geographic information online. Allied to this is the increased crowd sourcing by online communities of *volunteered geographic information* (VGI; Goodchild, 2007) and *user-generated content* (UGC). As such, neogeography is founded upon the two-way, many-to-many interactions between users and websites that have emerged under Web 2.0, as embodied in projects such as Wikimapia (www.wikimapia.org) and OpenStreetMap (www.openstreetmap.org). Today, Wikimapia contains user-generated entries for more places than are available in any official list of place names, and the term *vernacular region* is used to describe regions which emerge from geocomputational analysis of feeds from social networking sites. OpenStreetMap is well on the way to creating a free-to-use global map database through assimilation of digitized satellite photographs with GPS tracks supplied by volunteers.

This has converted many new users to the benefits of creating, sharing, and using geographic information, often through ad hoc collectives and interest groups. Such sites go some way to alleviating concerns about the social implications of GIS, insofar as participation in the creation and use of GIS databases is not restricted, and the contested nature of place names and other characteristics can be tagged in publicly editable databases. As such, Web 2.0 simultaneously facilitates crowd sourcing of VGI while making basic GIS functions increasingly accessible to an ever-broader community of users. This creation, maintenance, and distribution of databases has been described as a *wikification of GIS* (Sui 2008).

Official data are also becoming available through renewed pressures for government accountability, and the broader realization that wide availability of data collected by government and pertaining to citizens can lubricate economic growth. The result has been a plethora of open-data initiatives in many developed countries, leading to Web-based dissemination of data relating to many areas of public concern, such as personal health, transport, property prices, and even the weather. Conventional official sources such as censuses of population today account for a very much smaller proportion of the data that are collected about citizens, and there is a sense in which open-data initiatives are playing catch-up – providing researchers and analysts with some facility with which to understand the increasingly diverse and complex social, economic, and demographic milieu that characterizes advanced societies. Despite the hubris that has been generated around open-data initiatives, however, most of the data sources that have been released present extremely partial and disconnected representations of the world. For reasons set out in the discussion of modifiable areal unit effects above, the much more holistic concerns with issues of choice and service delivery, or the localism agenda in general, require linked characteristics at the level of the individual citizen or at the very least small neighborhood units.

This will require clear thinking of issues of spatial resolution (level of detail) and disclosure control that are central to the wider spatial literacy agenda (Janelle and Goodchild 2011). One consideration that is likely to reignite aspects of the social critique of GIS is that it is unlikely that privacy strictures can ever be absolute. Open-data initiatives are creating the need for a broader policy framework for data that responds to concerns of citizen privacy and confidentiality while remaining cognizant of the benefits that can accrue through opening up, integrating, and using the contents of government data silos. What level of data degradation is an informed public likely to be happy with, if it can be shown to bring benefits in terms of efficient and effective provision of public and private goods?

A related challenge is that empowerment of the many to perform basic (and even advanced) GIS operations brings new challenges to ensure that tools are used efficiently, effectively, and safely. Whether using official statistics or VGI, Web 2.0 can never be more than a partial and technological substitute for understanding of the core organizing principles and concepts of GIScience. These highlight the need to know and specify the basis of inference from the partial representations that are used in GIS to the world at large, yet such information is conspicuous by its absence from many VGI sources.

56.4 Conclusion

In undertaking a wide-ranging review of the achievements of GIScience and geocomputation, this chapter has also set out the principal issues and challenges that face these fields today. Improved computation and the facility to create, concatenate, and conflate large data sets will undoubtedly guide the future trajectories of the fields in the short to medium term. Ultimately, though, our focus in this chapter has been upon changes in scientific practice that may appear mundane but are nonetheless profound and far reaching. Good science is relative to what we have now, and improved understanding of data and their provenance is a necessary precursor to better analysis of spatial distributions in today's data- and computation-rich world.

Ultimately, GIScience and geocomputation are applied sciences of the real world and in large part will be judged upon the success of their applications. Improved methods and techniques can certainly help, as can ever-greater processing power. Yet the experience of the last 20 years suggests that there are rather few purely technical solutions to substantial real-world problems. The broader challenge is to address the ontologies that govern our conception of real-world phenomena and to undertake robust appraisal of the provenance of data that are used to represent the world using GIS.

This argues that the practice of GIScience and geocomputation poses fundamental empirical questions that require place or context to be understood as much more than location. Scientific approaches to representing places will undoubtedly benefit from the availability of new data sources and novel applications of existing ones, as well as citizen participation in their creation and maintenance. Yet a further quest for GIScience is to develop explicitly geographical representations of the accumulated effects of historical and cultural processes upon unique places.

References

- Adnan M, Longley PA, Singleton AD (in press) Parallel processing architectures of GPU: Applications in geocomputational geodemographics. In Abrahart R, See L (ed) *GeoComputation*, 2nd edn. Taylor and Francis, London
- Anselin L (1989) What is special about spatial data? Alternative perspectives on spatial data analysis. Technical Paper 89-4. National Center for Geographic Information and Analysis, Santa Barbara
- Anselin L (1995) Local indicators of spatial association – LISA. *Geograph Anal* 27(2):93–115
- Batty MJ, Longley PA (1994) Fractal cities: a geometry of form and function. Academic, San Diego
- Boucher A, Kyriakidis PC (2006) Super-resolution land cover mapping with indicator geostatistics. *Remote Sens Environ* 104(3):264–282
- Clarke KC, Gaydos L (1998) Loose coupling a cellular automaton model and GIS: long-term growth prediction for San Francisco and Washington/Baltimore. *Int J Geograph Inform Sci* 12(7):699–714
- Cope M, Elwood S (2009) Qualitative GIS: a mixed methods approach. Sage, Thousand Oaks

- Duckham M, Goodchild MF, Worboys MF (2003) Foundations of geographic information science. Taylor and Francis, New York
- Egenhofer MJ, Franzosa RD (1991) Point-set topological spatial relations. *Int J Geograph Inform Sci* 5(2):161–174
- Evans TP, Kelley H (2004) Multi-scale analysis of a household level agent-based model of landcover change. *J Environ Manage* 72(1–2):57–72
- Fotheringham AS, Brunsdon C, Charlton M (2002) Geographically weighted regression: the analysis of spatially varying relationships. Wiley, Hoboken
- Goodchild MF (1992) Geographical information science. *Int J Geograph Inform Sys* 6(1):31–45
- Goodchild MF (2007) Citizens as sensors: the world of volunteered geography. *GeoJournal* 69(4):211–221
- Goodchild MF (2010) Twenty years of progress: GIScience in 2010. *J Spat Inf Sci* 1(1):3–20
- Hey AJG, Tansley S, Tolle KM (2009) The fourth paradigm: data-intensive scientific discovery. Microsoft Research, Redmond
- Janelle DG, Goodchild MF (2011) Concepts, principles, tools, and challenges in spatially integrated social science. In: Nyerges TL, McMaster R, Couclelis H (eds) *The SAGE handbook of GIS and society*. Sage, Thousand Oaks, pp 27–45
- King G (1997) A solution to the ecological inference problem: reconstructing individual behavior from aggregate data. Princeton University Press, Princeton
- LeSage J, Pace RK (2009) Introduction to spatial econometrics. CRC Press, Boca Raton/London/New York
- Longley PA, Goodchild MF, Maguire DJ, Rhind DW (2011) Geographic information systems and science, 3rd edn. Wiley, Hoboken
- Maguire DJ, Longley PA (2005) The emergence of geoportals and their role in spatial data infrastructures. *Comp Environ Urban Syst* 29(1):3–14
- Maguire DJ, Batty MJ, Goodchild MF (eds) (2005) GIS, spatial analysis, and modeling. ESRI Press, Redlands
- National Science Foundation (2003) Report of the blue-ribbon advisory panel on cyberinfrastructure. National Science Foundation, Washington, DC
- Openshaw S (1983) The modifiable areal unit problem. GeoBooks, Norwich
- Openshaw S, Abrahart RJ (1996) Geocomputation. In: Abrahart RJ (ed) *Proceedings, first international conference on geocomputation*. University of Leeds, Leeds, pp 665–666
- Ord JK, Getis A (1995) Local spatial autocorrelation statistics: distributional issues and applications. *Geograph Anal* 27(4):286–306
- Parker DC, Manson SM, Janssen MA, Hoffmann MJ, Deadman P (2003) Multi-agent systems for the simulation of land-use and land-cover change: a review. *Ann Assoc Am Geograph* 93(2):314–337
- Pickles J (ed) (1995) Ground truth: the social implications of geographic information systems. Guilford, New York
- Shreve RL (1966) Statistical law of stream numbers. *J Geol* 74:17–37
- Sui D (2008) The wikification of GIS and its consequences: or Angelina Jolie's new tattoo and the future of GIS. *Comp Environ Urban Syst* 32:1–5
- Tobler WR (1970) A computer movie simulating urban growth in the Detroit region. *Econ Geogr* 46(2):234–240
- Torrens PM, Li X, Griffin WA (2011) Building agent-based walking models by machine learning on diverse databases of space-time trajectory samples. *Trans Geogr Inform Sci* 15(s1):67–94
- Wilson AG (1970) Entropy in urban and regional modelling. Pion, London
- Wright DJ, Goodchild MF, Proctor JD (1997) Demystifying the persistent ambiguity of GIS as “tool” versus “science”. *Ann Assoc Am Geogr* 87(2):346–362

Geospatial Analysis and Geocomputation: Concepts and Modeling Tools

57

Michael de Smith

Contents

57.1	Introduction	1123
57.2	Geocomputation and Spatial Analysis	1125
57.3	Geocomputational Models Inspired by Biological Analogies	1127
57.4	Networks, Tracks, and Distance Computation	1131
57.5	Computational Spatial Statistics	1132
57.6	Conclusions	1134
	References	1135

Abstract

This chapter provides an introduction to geocomputation and geocomputational methods. As such it considers the scope of the term geocomputation, the principal techniques that are applied, and some of the key underlying principles and issues. Chapters elsewhere in this major reference work examine many of these ideas and methods in greater detail. In this connection it is reasonable to ask whether all of modern spatial analysis is inherently geocomputational; the answer is without doubt “no,” but its growing importance in the development of new forms of spatial analysis, in exploration of the behavior and dynamics of complex systems, in the analysis of large datasets, in optimization problems, and in model validation remains indisputable.

M. de Smith

Department of Geography, University College London, London, UK

e-mail: mike@desmith.com

57.1 Introduction

For many researchers the term *geocomputation* refers to “the art and science of solving complex spatial problems with computers.” This definition, which is the tag line of the academic conference series run under the banner “geocomputation.org,” captures the essence of the term geocomputation. As such it embraces all manner of concepts, tools, and techniques that form part of mainstream geographical information systems (GIS) and science (GIScience: Goodchild and Longley, this volume*) and the methods employed in spatial analysis. For example, the 2011 Geocomputation conference included, among other topics, sessions entitled Geodemographics, Genetic Algorithms and Cellular Automata Modeling, Agent-Based Modeling (ABM), Geostatistics, Space-Time Modeling and Analysis, Network Complexity, Machine Learning, GeoVisual and Terrain Analysis, and Geographically Weighted Regression. Thus, parts of the academic world currently use the term geocomputation to apply to a very wide range of spatial analysis and modeling procedures, particularly those for which computational resources are central to the techniques employed. Readers will have noticed that each of the above topics, with the exception of machine learning, is represented by one or more chapters in this major reference work.

As Fischer and Leung (2001), Prologue have noted, geocomputation may also be viewed as a *research paradigm* that has changed the view of research practice in geospatial science over the past two decades. The driving forces behind the paradigm are fourfold: first, the increasing complexity of our spatiotemporal systems (nonlinearity, uncertainty, discontinuity, self-organization and continual adaptation); second, the need to develop new ways of utilizing and handling the increasingly large amounts of spatial information from the GIS and remote-sensing revolutions; third, the availability of attractive computational (intelligence) technologies which provide the modeling tools; and finally, the advent of high-performance computers.

As individual analytical methods that are considered geocomputational become accepted as providing effective solutions to specific spatial analysis problems, so they start to appear in more generic software with widespread usage. This is particularly apparent in some areas of remote-sensing data analysis, visualization tools, and agent-based modeling and in association with a number of statistical and spatial optimization problems. Thus, we are led to the conclusion that geocomputation is often used as an umbrella term for approaches to the analysis of problems that have a specifically geographical or environmental data focus, use modern computational techniques, and leverage high-performance computing hardware. This remains consistent with Openshaw’s original vision for the discipline (Openshaw and Abrahart 2000, p. x). As such, therefore, geocomputation is not a fixed set of techniques and models but an evolving, shifting collection of ideas, computational tools, and techniques with a particular emphasis upon the spatial domain, at scales from the architectural to the global. Modern geocomputational research remains true to the objectives Openshaw laid out more than a decade ago as being “...all about the use of relatively massive computation to tackle grand challenge (viz. almost impossible to solve) [geo]problems of immense complexity” (Openshaw and Abrahart 2000, p. 9).

In recent years many fields of scientific research have become dependent upon computational rather than so-called analytical methods. This is particularly apparent in mathematics, in areas such as combinatorial analysis, network analysis, and much of modern statistics. Likewise, developments in GIS software and remote sensing have relied very heavily on advances in computer memory, disks, processor speed, processing architectures, and visual display technology in order to provide the tools that are regarded as essential for digital mapping and related tasks. The challenge for geocomputational techniques is to provide real added value by providing practical tools that help solve complex problems, extract essential information from large datasets, and enhance the understanding of spatial processes and outcomes. Such techniques and modeling exercises may not be simple or parsimonious, but frequently they have the merit of providing more meaningful outcomes that match real-world experiences more closely than traditional analytical models. The real world (physical and social) is intrinsically complex, dynamic, and often unpredictable, and thus, computational spatial modeling tools that embrace this complexity can make a powerful contribution to our understanding of how the world works.

57.2 Geocomputation and Spatial Analysis

In our introductory section, we noted that geocomputation is a term that is applied to computationally demanding methods of spatial analysis. By definition problems that are addressed by geocomputational methods push the boundaries of computing technology. Since this is a rapidly changing field, with continuous improvements in processing power, memory availability and speed, networking, and storage capacity, yesterday's geocomputational problem may well become today's standard procedure. Problems that had previously to be run on limited size datasets, or with severe limitations to the spatial and temporal resolutions employed, can in due course be run with much finer resolution datasets, with more realistic assumptions, and with many repetitions if so required.

Specialized computing architectures can be leveraged very effectively for many spatial problems – for an excellent recent discussion of this field, see Huang et al. (2011). For example, it is becoming straightforward to use multiple processors on single computers or across multiple computers (e.g., using grid networks or cloud computing: Adnan et al. this volume*) to run many simulations in parallel that are identical other than with respect to their initial and boundary conditions. Likewise, classification techniques, such as k-means clustering, can be readily implemented using parallel systems processing since the procedure involves a series of similar runs for $k = 2 \dots$ to 20 say, iterated for a set of (randomly) chosen initial seed locations. In other instances the software application itself requires substantial redesign in order to operate effectively across multiple processors and memory regions. This is particularly true for data-intensive applications, where the dimensionality of the problem may be limited but the volumes extremely large (e.g., very high-resolution satellite imagery and LiDAR (Light Detection And Ranging) point

clouds, determination of optimal parameters for fine-resolution spatial interaction models). In such cases splitting the problem into smaller units (e.g., decomposition of the data into tiles or cells), processing the separate units in parallel, and then merging the results may be the only practical approach that can be adopted.

With the rise of increasingly large spatial datasets, almost all aspects of spatial data processing become a challenge, but with geocomputation the focus is upon problems that resist classical forms of analysis. One of the clearest examples of this arises in the field of spatial simulation (often now referred to as geosimulation – see the work of Paul Torrens at <http://geosimulation.org/>). Although there are a variety of approaches to such simulation – for example, cellular automata, agent-based modeling, and randomized network-based modeling – the results obtained at the macroscale from microscale (or bottom-up) modeling are often unexpected and unpredictable. Macroscale structures and behaviors “emerge” from simple microscale processes, in a similar manner to the emergence of macroscale features in biological systems – such as the appearance of segmented body structures and articulated limbs in a wide range of insects and higher life forms and the appearance of “swarms” or flocking behavior among birds, fish, and many other animals. In geospatial analysis, examples include modeling the behavior of pedestrians during evacuation emergencies, examining the way in which disease spreads among communities, and modeling and predicting urban land use and transport changes over time.

Microscale or “bottom-up” simulation is only one of a number of major application areas for geocomputational methods. Other examples include computational spatial statistics – procedures that seek to provide statistically valid insights into complex spatial datasets; optimization problems, ranging from optimal location and routing problems to the determination of optimal model parameters for highly parameterized models; procedures that seek to augment data at a given scale and point in time with related datasets at different levels of aggregation and/or extending over a number of time periods; and a wide range of advanced visualization techniques. In many instances these issues are not separate, independent concerns but apply simultaneously to the iterative process of model building, thereby demanding considerable skill in model construction, software engineering, data management, and validation. The results of such work are often surprising and impressive but may also be difficult to follow in detail and hard to justify. Is understanding always advanced when such methods are employed or do such models match the past and present so well because they are highly parameterized and intensively fitted?

In the sections that follow, we commence in Sect. 57.3 by examining a number of geocomputational methods that have been inspired by analogy with biological processes. These include cellular automata and agent-based models, computational neural networks, and evolutionary algorithms. In Sect. 57.4 we then discuss a number of geocomputational techniques that have been applied to network-related problems rather than point- or areal-based spatial problems. In Sect. 57.5 we look at the rising importance of computational methods in statistical science and the impact of this development to the field of spatial statistics. We conclude in Sect. 57.6 by commenting upon issues of spatial and temporal resolutions and questions of model complexity.

57.3 Geocomputational Models Inspired by Biological Analogies

The analogy with biological processes runs far deeper in the field of geocomputation than might be expected. Three major classes of methods commonly applied in geocomputation have been inspired by biological analogies (see, further, de Smith et al. 2009, Chap. 8). The first of these embraces a number of simulation techniques, many of which owe their origins to a simple cellular automaton (CA) simulation known as “the Game of Life.” This extremely simple model, which operates within the framework of a 2D matrix of square cells, was introduced by the mathematician John Conway in the 1970s and has a surprisingly extensive and complex range of outcomes. The rules of the Game can be summarized as follows:

The Game of Life has only two *states* for each cell (alive or dead, or 1 or 0) and three simple *state transition rules*: (i) *survival*: if a cell is ‘alive’ (i.e., its *state* is ‘true’ or ‘1’, for instance) and it has two or three alive neighbors it remains alive; (ii) *reproduction*: if a cell is ‘dead’, but has three alive cells within its *neighborhood* its *state* becomes alive, and (iii) *loneliness* (less than two neighbors) or *overcrowding* (more than three neighbors): the cell dies. Despite these simple rules, the Game of Life can produce a range of complex behaviors from different initial conditions.

This archetypal spatial model demonstrates an important characteristic of many geocomputational methods – the emergence of broad classes of pattern that could not have been predicted prior to simulation. It also highlights a number of other common features of such procedures: the outcomes of geosimulation exercises have an element of framework dependency (structures and boundaries) and, in some instances, path dependency (i.e., the future path is heavily determined by past steps and initial conditions and may lead to stable or unstable/widely divergent results). Hence, key structural attributes of each type of simulation model have a direct bearing on the kinds of outcome seen. These attributes (for problems of the cellular automata type) include state variables (the possible states the system supports and the set of initial states that are examined); the spatial framework used for modeling (e.g., use of a 2D rectangular grid, a 3D lattice, a toroidal space, etc.); the form of neighborhood effects permitted (e.g., single or multiple steps, distance bands, different forms of adjacency); the state transition rules applied (i.e., how the states may change over time); and finally, how the time dimension is treated – discrete or continuous, serial or parallel.

Wolfram (1983) studied the Game of Life in detail and demonstrated that the outcomes for all simulations of this general type fall into four classes. For the Game of Life, these are (i) static patterns, (ii) oscillators (or periodic patterns), (iii) spaceships (patterns that repeat themselves but translated in space), and (iv) patterns that increase in population size (a range of different patterns and behaviors). A useful summary with examples and references is available on the Wikipedia page “Conway’s Game of Life.” The unexpected complexity of the resulting behavior from the rules Wolfram identified led him to believe that complexity in nature may be due to similar behavior.

Batty (2000) provides an excellent review of CA models and their application to urban systems modeling, including commentary on their strengths and weaknesses, and identifies areas for future research in this field. Central among these are questions of improved representation of the real world, moving from the fixed framework of square cells to incorporate irregular zones, streets, and other linear forms, and new definitions of how such elements interact. To an extent, these issues have been tackled in recent years using an alternative geosimulation framework, known as agent-based models (ABMs).

Commencing with this simple cellular-based simulation framework, computer scientists have developed a large and expanding family of techniques and tools based on microscale or bottom-up simulation. Among these are a range of approaches that liberate the simulation from the spatial constraints of a lattice-based framework and permit-free movement in 2D or 3D space (agent-based models). One example is the so-called ant colony optimization (or ACO) approach, in which the space of interest is explored by synthetic ants (the agents in this case) – those that reach a desirable objective (e.g., a target location/food source) by any route then return to the ant colony by a relatively direct route laying a synthetic pheromone trail. Subsequent ants are attracted to the pheromone trail reenforcing its usage. But because the pheromone evaporates over time, only the route or routes that are most used tend to be retained, and since shorter routes lead to and from the food source more quickly, these tend to be reenforced at the expense of longer (i.e., less desirable) routes. ACOs are a form of agent-based model (where the ants are the agents), and because of the way such systems are implemented, with large populations of independent agents acting in a collective manner, ACOs and similar procedures fit within the broad area known as “swarm intelligence.” Other microsimulation procedures within this broader paradigm have been applied with considerable success to crowd behavior modeling, for example, helping to manage large-scale street events such as carnivals and protest marches, football stadiums, and emergency evacuation of complex buildings. Note that in these more advanced geocomputational models, a clear distinction exists between the population (agents in this case) and the environment. This is not the case with cellular automata. There is also the inherent assumption that while the population responds to the environment, the reverse is generally not the case – for some applications (e.g., those involving anthropogenic change), such assumptions are not realistic.

Recently a number of urban geosimulation models have sought to reflect real-world spatial structures and their attributes within the model framework. In these models the spatial framework for the model utilizes fine-scale digital maps of the study area combined with synthetically generated population data. A synthetic population is a computer-generated randomized set of individuals, households, or other entities of interest that match predefined aggregate attributes (or classifications) for zones within a study region. Such populations are often used as the building blocks for microsimulation projects of the kind described above, where the necessary individual-level data does not exist and a match to “ground truth” information is required (e.g., in land use planning, a traffic simulation model, a medical study, or even a model of burglary events that incorporates the potential

victims of crime). Typically the finest level of available zonal data (such as small area census zones) is used to specify the attributes of the synthetic population. The aggregated characteristics of this computer-generated population are chosen to ensure that they closely approximate the zonal figures. A recent comparison of procedures for achieving such approximations is provided by Heppenstall et al. (2011), whose tests identify the use of simulated annealing (SA) as the most effective (if relatively slow) procedure for minimizing classification error. The synthetic individuals are then allocated to particular locations within the study area zones. This might be a process of random allocation to known buildings or by reference to land use maps. In many instances the most practical option is to rasterize the zones and to allocate the synthetic population to cells based on a probabilistic assignment derived from land use. The result is a representative population of individuals, allocated to meaningful point or cell locations (thereby minimizing spatial aggregation bias) that may be used within a microsimulation model. This lends such models credibility and a level of realism derived from the knowledge that their aggregate characteristics match those that are known, even though they do not strictly represent the actual set of individuals within the study area. Thus, a form of “study error” remains, but the approach goes some way to resolving the ever present issues of statistical and spatial aggregation associated with many datasets.

The second main area of biologically derived geocomputational methods is the field known as computational neural networks (CNNs). These methods take their inspiration from highly simplified models of neurological processes, originally drawing on models of how neurons in the retina operate. In practice the analogy has proved useful in guiding the general structure of computational models rather than any direct appeal to biological process similarity. CNN methods have been successfully applied to a limited number of geospatial problems – most notably the modeling of trip distribution data (see, e.g., Fischer 2006) – and for multispectral image analysis. Typically three layer models have been applied, with an input layer, single middle layer, and an output layer. CNNs are models whose behavior is determined in part by the model structure itself and in part by the data used to “train” the model parameters. Once the structure has been specified by the research scientist (which may be an iterative process in itself) and training data selected, applied, and evaluated, the result is a modeling system that can be used on more general “unseen” datasets. As with the previous discussion of systems inspired by artificial life, the outcomes achieved are strongly dependent on a number of well-defined attributes of the model: the number of layers used in the neural model, the number of nodes used in each layer, the form of the forward and backward propagation algorithms, and the connectivity of the network. This dependency on the structure of the model is then accentuated by an element of data dependency since the selection of the training dataset and control (evaluation) dataset defines the parameters that are then used on unseen data. CNNs have found their most widespread geospatial application in the field of remote sensing, where they have been found to be a very effective tool for inferring land use and land use change. However, the nature of such models suggest that they operate primarily as a form of pattern recognition engine, a relatively highly

parameterized black box, that like higher animals can be very good at recognizing familiar shapes and textures but by itself contributes little to understanding and thus can be quite specific to the datasets any specific model can be applied to. In the longer term a combination of high speed automated CNNs with “intelligent systems” may produce much more generic and flexible tools that offer a real step forward in efficient processing of large spatial datasets.

The third biological analogy we discuss here is “survival of the fittest.” Here the objective is to solve a range of difficult combinatorial optimization problems using analogies from genetics. There is a wide range of important optimization problems for which it is believed no solution algorithm exists that can be run in polynomial time – that is, such problems are intrinsically nonpolynomial (or NP). What this means in practice is that for problems that involve any realistic quantities of data (i.e., n objects, where n is not small), it is impossible to find a provably globally optimal solution in a finite amount of time. As n increases so the problem becomes increasingly intractable. The archetypal example of such a problem is the simple traveling salesman problem (TSP). The TSP involves finding the optimal route through a set of points (e.g., towns) that ensures each is visited just once and the total length of the tour is minimized. The best algorithms for solving the simple TSP can now provide provably optimal solutions for n in the 1,000 s, which is very impressive. Genetic algorithms (GAs) can be used to “solve” the TSP but are far less efficient than many other approaches – that is, GAs typically produce poorer results (suboptimal solutions) and take longer to achieve these (many 1,000 s of iterations or “generations”). One reason for this is that GAs are a generic rather than specific approach to problem solving – they can often provide a reasonably good answer rather than the best answer. And this raises another interesting question – is the optimal solution the best solution? Initially we would immediately respond “of course,” but in many instances this perspective must be qualified by recognizing that it applies to the problem as specified, which is typically very tightly specified and generally static. If the kinds of problem to be tackled are less well specified and dynamic or involve more complex choice models, it is possible that apparently suboptimal computational methods might well fare far better than fixed algorithms.

GAs are a particular subset of a broader class of algorithms known as evolutionary algorithms (EAs). EAs have been applied to a variety of dynamic optimization problems with some success. This raises the question of what is meant by “dynamic.” Clearly all problems exist in a temporal context but are only considered dynamic if some aspect of the problem changes over time. Key issues are therefore the frequency and scale of any such changes, their behavior (e.g., predictable or random, trending or cyclical, or some combination of these), and the nature of the changes taking place – are the relevant dynamic elements changes in the environment the objective function or perhaps constraints that affect the system behavior? In a static optimization problem, a traditional genetic algorithm attempts to converge toward a global optimum by a process of genetic selection – retaining genetic strings that are fitter (provide solutions closer to the optimum, as measured by some fitness function) and modifying genes by various forms of mutation and crossover. Since these strategies are designed to migrate the solution toward a static optimum, they are

too restrictive for problems that include dynamics. Dynamic optimization problems require modification of such survival strategies by taking into account the nature of the changes that may occur. In some instances such changes are so dramatic and unpredictable that attempts at developing solution procedures will never succeed, but in many instances changes are smoother or have a more predictable effect (e.g., reducing the severity of some constraints), and in these cases dynamic EA procedures can be used. Example approaches that have been found to be effective include the concept of introducing random immigrants into the population, helping to maintain genetic diversity, and so-called forking of populations, in which an entire subset (e.g., a selection of children) are separated off and explore their own evolutionary path transposed from the parent population (a form of emigration).

57.4 Networks, Tracks, and Distance Computation

Many of the techniques described above apply to point and areal datasets (zonal or lattice-based), although some relate to networks. Until recently attention on linear forms has tended to concentrate on a range of network optimization problems, such as determining shortest paths, least cost routes, multivehicle (capacity constrained) route allocation, optimal on-network facility location, and optimal arc routing (e.g., street cleaning) – see de Smith et al. (2009), [Chap. 7](#) for a detailed review of this area. In many cases very fast algorithms have been obtained to solve such problems, and in some cases these solutions are known to be optimal. In other cases achieving a provably optimal result may be impossible and/or require unlimited computed resources for large problems (i.e., problems with many links and nodes). As a result a wide range of suboptimal but effective procedures have been developed, which provide very acceptable if not optimal solutions. Among these are highly problem-specific algorithms (e.g., the A* algorithm for determination of the shortest path in a network), various forms of linear programming and dynamic programming, and much more generic procedures, such as genetic algorithms (GAs), cellular neural networks (CNNs), simulated annealing (SAs) algorithms, ant colony optimization (ACO) methods and many more.

In most cases these models and methods seek optimal or near optimal solutions to complex routing problems in static network environments. However, it is increasingly apparent that solution procedures need to reflect the dynamics of real-world networks, where road traffic or telecommunication traffic intensity and connectivity may vary rapidly in time and space (see Cheng, this volume*). Thus, geocomputational systems need to be able to respond in near real time to events as they occur, particularly when applied in command and control situations.

The emergence of new forms of data flows, most notably from individuals (people, animals) or vehicles on the move and from satellite tracking and sensing, has led to a substantial new body of spatial data that demands our attention. In some instances these data need to be fed directly into the processing systems in order to identify problems and direct changes to the way these systems manage and forecast near-term events (e.g., traffic routing, crowd control, evacuation management), while in other

cases the data can be seen as a complex series of tracks, marking out the routes chosen by numerous individuals or other entities over time. This tracking information, often generated by communications devices that incorporate Global Positioning System (GPS) or similar technologies, provides an entirely new body of complex spatial data that demands new forms of analysis to interpret and leverage results. Research into the use of such data is at an early stage, with initial applications in fields such as attempting to predict animal behavior (e.g., migration routes, preferred habitat selection), modeling the behavior of would-be burglars, and analyzing the tracks of hurricanes.

Accurate computation of distance is a fundamental requirement of almost every form of geocomputation, but in many instances approximations are used that may result in systematic errors. The main issues are the use of Euclidean measure when network-based measures are more appropriate and the use of local Euclidean distances when undertaking operations across raster files (see further below). These issues become more complex when constraints are added, for example, when restricted areas are included, barriers and turn restrictions are accounted for, and where the underlying space is no longer treated as being homogeneous.

In the first case, Euclidean measure may substantially underestimate interpoint distances. Computation of network distances across dense street networks can be time consuming, with the result that some GIS operations such as calculating the extent of drive-time polygons, or computing optimal facility locations, may be very slow. Where distance computations are not based on networks and are computed using Euclidean, spherical, or geodesic measure, they can be calculated extremely quickly. In some instances such measures may be used as a surrogate for network distance or as a means of generating good “candidate lists” of solutions which can then be used as initializations for network-based optimization.

In the second case, distance computations on raster files, many applications use operations on the immediate eight neighbors of each cell in the raster (i.e., work on a 3x3 array of cells). Where these involve distance calculations, for example, when the lengths of paths are computed across a cellular model or hydrological flows analyzed in a digital terrain model, the distance computations are often incorrect. Local Euclidean distances (1 for N, S, E, W movements and $\sqrt{2}$ for diagonal compass movements) result in errors of almost 8 % in overall distances and result in errors in optimal path selection. The solution in such cases is to use optimal 3x3 or 5x5 values (real or integer approximations) or to use exact Euclidean distance transforms (DTs) that correct the propagated errors (see, further, de Smith et al. 2009, Sect. 4.4.2.2). Furthermore, DTs may be used as a geocomputational procedure that can solve a variety of spatial optimization problems, notably optimal route finding in nonuniform free space (nonnetwork) environments where gradient, curvature, and other constraints apply.

57.5 Computational Spatial Statistics

Our earlier section on biologically inspired models embraces a large part of the subject we have been referring to as geocomputation. To some extent in parallel,

a separate and rather different area of geocomputation has been developed. This involves the application of computationally intensive methods to a range of problems arising in spatial statistics. Openshaw's (1987) groundbreaking work in this field involved attempting to apply raw computing power to identify potentially significant clusters of point-referenced data such as the incidence of rare diseases (e.g., cases of childhood leukemia).

More recently geocomputational methods have been developed that provide similar functionality within a more statistically robust framework. Principal among those now widely used are spatial scan statistics, originally developed by Kulldorf (1997) at the National Cancer Institute in the USA. Kulldorf's scan procedures have been developed and extended over the years and many of these developments have been implemented in the software package SaTScan (for a recent example application, see Greene et al. 2010).

As the SaTScan authors' state, the software is designed to:

- Perform geographical surveillance of disease, to detect spatial or space-time disease clusters, and to see if they are statistically significant
- Test whether a disease is randomly distributed over space, over time, or over space and time
- Evaluate the statistical significance of disease cluster alarms
- Perform repeated time-periodic disease surveillance for early detection of disease outbreaks

This description of how SaTScan operates, as a computationally intensive approach to spatial statistical analysis, is typical of many forms of modern spatial analysis. It utilizes computational power to search for and examine patterns within large volumes of data. In this case the procedures applied involve a scanning process, similar in concept to scanning algorithms applied in remote sensing and image processing (e.g., as used in the computation of distance transforms, noted above) – it is fast, simple, and exhaustive. In many other instances the search space is more complex, often multidimensional, and the search procedures cannot be exhaustive but must rely on heuristics, including those biologically inspired procedures described earlier.

Another feature common to geocomputational statistics is the production of pseudoprobability distributions by large-scale simulations or random permutations. These procedures recognize the limitations of traditional analytical methods and seek to use observations and/or aspects of observed spatial structure to generate a large set of possible outcomes under conditions of randomization. Using these simulations one or more statistics of interest are computed and the observed value in a given dataset is then regarded as a particular realization within the simulated population. If the observed statistic appears exceptional (e.g., the mean distance to nearest neighbor in a bounded point set is measured and found to be very small or very large when compared to a large number of random simulations using the same number of points within the same spatial extent), then it can be regarded as "significant." Care must be taken when carrying out such procedures to avoid generating pseudoprobability distributions that involve resampling the same data (or regions) multiple times and to take into consideration questions of the independence of samples.

Likewise, for data relating to planar lattices (e.g., all census tracts within a state), statistics may be computed and compared with a large number of permutations of the observed data values across the lattice. This procedure, which is widely used to generate pseudoprobability distributions and evaluate the “significance” of observed patterns, is less satisfactory from a statistical perspective. It assumes that the data values for the M zones are effectively fixed and could have been allocated at random to any of the M zones in the study area. In practice this is not the case – it is more reasonable to assume that the total count (e.g., of cases of a specific illness) could have been randomly divided into M partitions (based on a uniform or observed frequency distribution) and each then assigned at random to one of the zones. The assigned values may then be used to calculate the statistics of interest, with repetition of this process yielding a pseudoprobability distribution that can then be used for comparative purposes. However, this procedure also has limitations since the partitioning and allocation among zones is purely random and may not reflect important variations between these zones, for example, in terms of the size of the population at risk. Furthermore, establishing an appropriate null hypothesis in such cases may be difficult or impossible since zonal boundaries are often arbitrary and some level of spatial autocorrelation is always present.

Procedures such as geographically weighted regression and spatial regression models (see, further, Spatial Econometrics, this major reference work*), spatial analysis on networks (SANET), and geostatistical modeling can also be considered as geocomputational methods. They rely on computational power to produce insights into the statistical significance of patterns and to facilitate model building where the parameter space is large and the observed datasets increasingly large and complex. Other computationally intensive statistical procedures, such as Markov Chain Monte Carlo (MCMC) techniques widely used in Bayesian model building (e.g., the GeoBUGS project), bootstrapping, and cross validation, all may be considered as having close links with the field of geocomputation.

57.6 Conclusions

A feature of many spatial datasets is their increasing complexity and size – many datasets are now available at a variety of levels of aggregation, not always relating to the same data. Techniques for maximizing the use of such data – for example, combining point and areal measures – are becoming more common. Likewise there is a growing availability of spatial datasets for different time periods, varying from relatively long period time slices (e.g., annual), to monthly, daily, or even shorter-time windows. This has led to researchers seeking to revise some traditional models, such as those relating to spatial interaction and trip distribution, to incorporate diurnal variations in population (e.g., home-based and work-based shopping trips, evaluating the optimal location of ambulances or police patrols). In such cases the objective is to improve the quality of models in order to provide improved understanding of the processes at work and the outcomes that can reasonably be predicted.

However, in a wide ranging discussion of computational model building, Batty (2011), p. 28 argues that “as the level of detail in terms of sectors, spatial-locational resolution, and temporal resolution increases, data demands generally increase and models become increasingly difficult to validate in terms of being able to match all the model hypotheses . . . to observed data. As temporal processes are added, this can become exceptionally difficult. . . and we face a severe problem of validation.... This tends to force modeling back to the traditional canons of scientific inquiry where parsimonious and simple models are the main goal of scientific explanation.” Thus, there is much to be gained from advances in model construction using computationally intensive procedures, but the results can be very difficult to validate, comprehend in detail, or justify to stakeholders, leading to considerable tension among the research community. Advances in visualization, ranging from improved 2D and 3D graphics to new metaphors for data exploration (e.g., Google Earth, dynamic fly throughs, immersive systems) and video display of the progress of geosimulations, all help to overcome some of the issues raised by Batty. And despite the difficulties, as the program for the 2011 Geocomputation conference identifies, this field is one attracting intense interest, is of great practical significance, and is set to become one of the defining scientific paradigms of the twenty-first century.

References

- Batty M (2000) Geocomputation using cellular automata. Ch 5. In: Openshaw S, Abrahart RJ (eds) Geocomputation. Taylor and Francis, London, pp. 95–126
- Batty M (2011) A generic framework for computational spatial modeling. Working Paper 166, Centre for Advanced Spatial Analysis (CASA), UCL, London. Available from <http://www.casa.ucl.ac.uk/publications/workingPaperDetail.asp?ID=166>. Accessed 4 Oct 2011
- de Smith MJ, Goodchild MF, Longley PA (2009) Geospatial analysis: a comprehensive guide to principles, techniques and software tools. 3rd ed. Troubador Publishing, Leicester. Also available online at <http://www.spatialanalysisonline.com>
- Fischer MM (2006) Spatial analysis and geocomputation: selected essays, vol 1. Springer, Heidelberg
- Fischer MM, Leung Y (eds) (2001) GeoComputational modelling. Techniques and applications. Springer, Berlin/Heidelberg/New York
- Greene SK, Schmidt MA, Slobierski MG, Wilson ML (2010) Spatio-temporal patterns of viral meningitis in Michigan, 1993–2001. In: Fischer MM, Getis A (eds) Handbook of applied spatial analysis. Software tools, methods and applications. Springer, Berlin/Heidelberg/New York, pp. 721–735
- Heppenstall AJ, Harland K, Smith DM, Birkin MH (2011) Creating realistic synthetic populations at varying spatial scales: a comparative critique of population synthesis techniques. Geocomputation 2011 Conference Proceedings, UCL, London, pp. 1–8
- Huang Q, Yang C, Li W, Wu H, Xie J, Cao Y (2011) Geoinformation computing platforms. Ch.3, pp. 79–126 in Yang R et al. Advanced Geoinformation Science. CRC Press, Baton Rouge, FL, op. cit
- Kulldorf M (1997) A spatial scan statistic. Commun Stat: Theor Method 26:1481–1496
- Openshaw S (1987) A mark 1 geographical analysis machine for the automated analysis of point data sets. Int J Geogr Inf Syst 1:335–358

- Openshaw S, Abrahart R J (eds.) (2000) Geocomputation. Taylor and Francis, London. Wikipedia: conway's game of life. Available from http://en.wikipedia.org/wiki/Conway%27s_Game_of_Life
- Wolfram S (1983) Statistical mechanics of cellular automata. Rev Mod Phys 55:601–643

Ross Maciejewski

Contents

58.1	Introduction	1137
58.2	Statistical Graphics	1138
58.2.1	Histograms	1139
58.2.2	Box Plots	1140
58.2.3	Scatterplots	1141
58.2.4	Parallel Coordinate Plots	1142
58.3	Choropleth Maps	1143
58.3.1	Color	1145
58.3.2	Class Intervals	1146
58.4	Exploratory Data Analysis	1147
58.5	Exploring Time	1150
58.5.1	Animation	1152
58.5.2	Space-Time Cube	1152
58.6	Conclusion	1153
	References	1154

58.1 Introduction

The current ubiquity of data collection is providing unprecedented opportunities for knowledge discovery and extraction. Data sources can be large, complex, heterogeneous, structured, and unstructured. In order to explore such data and exploit opportunities within the data deluge, tools and techniques are being developed to help data users generate hypotheses, explore data trends and ultimately develop insights and formulate narratives with their data. These tools often rely on visual representations of the data coupled with interactive computer interfaces to aid the

R. Maciejewski
Arizona State University, Tempe, USA
e-mail: rmacieje@asu.edu

exploration and analysis process. Such representations fall under the purview of visualization, in which scientists have worked on systematically exploiting the human visual system as a key part of data analysis. Research in this area has been inspired by a number of historical sources, examples include physicist James Maxwell's sculpture of a thermodynamic surface in 1874, Leonardo da Vinci's hand-drawn illustration of water from his studies to determine the processes underlying water flow, or the flow map of Napoleon's March on Moscow produced by Charles Minard in 1869. Each of these examples attempts to explain data in a visual manner, and, as visualization has progressed, principles and practices have been adopted to standardize representations, and, more importantly, better exploit properties of the human visual system.

In this chapter, we will focus on how visualization research has effectively utilized one of the most ubiquitous visual representations, the cartographic map. Cartography is the study and practice of making maps and is situated as perhaps the most well-studied visualization technique available to scientists. For centuries, cartographers have looked at combining science, aesthetics, and analysis into the mapmaking process on the premise that such tools will be able to effectively communicate information and aid in knowledge generation. One of the most famous examples of such knowledge generation is John Snow's mapping of the location of cholera deaths in Soho, England, in 1854. By plotting cholera deaths and the locations of water pumps, Snow was able to develop a hypothesis about the water-borne nature of the disease. Snow was able to use his visual explorations of the cholera outbreak patterns to persuade the local town council to disable the water pump central to the disease center.

This sort of data analysis utilizing cartographic principles as a means of representing spatiotemporal data and exploring patterns within this data is often referred to as *geographic visualization* or *geovisualization*. Geovisualization focuses on visually representing spatiotemporal data, exploiting known cartographic techniques as part of the interactive graphical representation of the data, and incorporating dynamic interactions for querying and exploring data. A rigorous definition of geovisualization can be found in MacEachren (1994): "Geographic visualization (can be defined) as the use of concrete visual representations - whether on paper or through computer displays or other media - to make spatial contexts and problems visible, so as to engage the most powerful human information-processing abilities, those associated with vision."

Note that the above definition discusses the use of visual representations (plural) of the data. Representations can range from complex glyphs, shaded areas, lines, and diagrams, and given the amount of data and differences in ways to represent data, it is important to consider what sort of questions will be asked of the data. Thus, rather than trying to make one "best" map or data representation which depicts only a subset of the available information, geovisualization systems often incorporate a variety of data views in an attempt to generate more insight for analysts. Often, the views generated involve the computation of basic statistics and summaries of the data as a means of providing the user with an overview of their data prior to, or as part of, the exploration process.

58.2 Statistical Graphics

As perhaps a precursor or overlapping field with data visualization, much of the groundwork in describing and exploring data sets comes from the *statistical graphics* community. Statistical graphics are tools used to reveal details about data sets, such as outliers and trends within the data. By revealing such features, one can speculate on how data can be further processed, transformed, explored, and analyzed. Hypothesis generation would begin in this stage, and such hypothesis generation would lead to the use of certain methods for hypothesis testing and data modeling. Finally, proper statistical methods could then be chosen for a further refined analysis of the data.

What is of key importance in data visualization is that when a statistical graphic is made, the information from the data set being explored is encoded by the chosen display method. An analyst will look at the visual representation, and then a decoding process will occur. During this decoding process, visual perception becomes the vital link. No matter how impressive the encoding is, if the visual decoding cannot be done by the analyst, then the data analysis will fail. This encoding and decoding of graphical elements is explored in *The Elements of Graphing Data* (Cleveland 1985), and detailed perceptual studies that discuss how well humans are able to perceive encodings (relative angles, line lengths, etc.) are described. Much use has been made of these studies in further developing visualizations, and a variety of methods are discussed in *Visualizing Data* (Cleveland 1993). Three of the most common exploratory data analysis graphics used for exploring the distribution of data include the box plot, the histogram, and the scatterplot (Fig. 58.1).

58.2.1 Histograms

One of the first things explored when analyzing data is the distribution of the given data measurements. According to Wilkinson (2005), the histogram (Fig. 58.1, left) is one of the most widely used visual representation and first-look analysis tool. Introduced by Pearson (1895), the histogram provides a visual summary of a univariate sample within a data set. The visual summary consists of rectangles drawn over discrete intervals (called *classes* or *bins*) where the height of each rectangle corresponds to the number of data samples that would fall into a given bin.

The main concern in creating a histogram lies within the choice of the number of bins and the width of the bins. Different numbers of bins and different bin widths can each reveal different insights into the data. Thus, the initial choice of the bin number and size can have a dramatic impact on the knowledge that can be derived when using a histogram visualization. Most statistical graphics programs default to one of two options when creating a histogram: the *square root choice* or *Sturges' choice* (Sturges 1926). The square root choice is defined such that given a data set with n samples, the number of bins k will be calculated as follows:

$$k = \lceil \sqrt{n} \rceil. \quad (58.1)$$

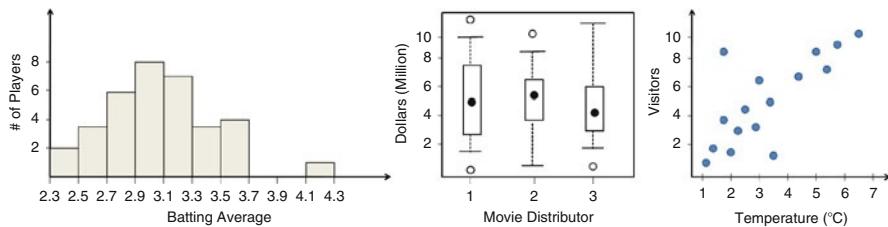


Fig. 58.1 Common statistical graphics. (*Left*) Histogram. (*Center*) Box plot. (*Right*) Scatterplot

Once the number of bins are known, it is assumed that each bin will be of equal width. Thus, the bin width h is defined as follows:

$$h = \left\lceil \frac{\max(x) - \min(x)}{k} \right\rceil \quad (58.2)$$

where x is the univariate data set under analysis. For comparison, Sturges' choice (Sturges 1926) is defined as follows:

$$k = \lceil \log_2(n) + 1 \rceil \quad (58.3)$$

Both the square root choice and Sturges' choice have an implicit assumption of a normally distributed data set. However, in the construction of the approximation, Sturges considered an idealized frequency in which the distribution will approach the shape of a normal distribution. Thus, if the data is not normal, the number of bins chosen will need to be revised. However, for moderately sized n (approximately $n < 200$), Sturges' rule will produce reasonable histograms. For further details on other histogram bin choices, please refer to (Wilkinson 2005).

While a vast amount of research has been done on bin selection, the key factor to take away is that there are strengths and weaknesses behind bin choices. If small values of h are used (with respect to the data in question), the histogram will model fine details within the data; however, noise in the data set can obstruct the information being presented and may obfuscate the analysis. Conversely, with large values of h , the density becomes too smoothed and one can oversimplify the description of the data.

58.2.2 Box Plots

While a histogram allows us to explore the distribution of a univariate measurement within our data set, an analyst often wants to quickly determine the mean values of the data distribution and compare this distribution to others. One measure that is key to visualizing data distributions is the *quantile*. The f quantile, $q(f)$, is defined as the value along the data measurement scale where approximately a fraction f of the data are less than or equal to $q(f)$. In the case of quartiles, this would be where

approximately one fourth of the data is less than $q(f)$ for the *lower quartile*, half for the second quartile (or the *median*), and three fourth for the *upper quartile*. The benefit of such measures is that the f-values provide a standard for comparisons across distributions.

One graphical way of using quantiles to compare distributions is the *Box plot*. Box plots, Fig. 58.1 (middle), consist of several distinct graphical elements, namely, the *box* and the *whiskers*. The box itself represents the range from the lower quartile to the upper quartile of the data, with the bottom edge of the box being the lower quartile and the top edge of the box being the upper quartile. The dot in the middle of the box represents the sample median. Note that if the dot is not directly in the center of the box, then it is an indication of data skewness.

The whiskers are the dashed lines that extend above and below the box. These lines represent the extent of the remaining data samples, and, assuming no outliers, the maximum value of the distribution is represented by the top line attached to the whisker, and the minimum value of the distribution is represented by the bottom line. If outliers exist, a small circle at the top or bottom of the plot will be used to represent this fact. A data sample is considered to be an outlier if its value is more than 1.5 times the interquartile range away from the top or bottom of the box.

While a box plot may seem simpler (or even less intuitive) than a histogram, its main advantage is its limited screen space requirements. A larger number of box plots can be plotted on the screen at once for comparison than would be possible for entire histograms. Furthermore, the visual representation of a histogram is highly influenced by the choice of bin width; there is no such limitation on the box plot.

58.2.3 Scatterplots

While the box plot allows users to compare and summarize distributions of like groups, analysts often wish to search for correlations between variables within their data set. One common means of visually representing the relationship between two variables within a data set is the *scatterplot*. Scatterplots visualize multi-dimensional data sets by assigning two of the data dimensions to a graphical axis, Fig. 58.1 (right). Points of the data set are then rendered in the Cartesian space defined by the axis. These plots are typically employed as a means of analyzing bivariate relationship within a planar projection of the data and are used to help researchers understand the potential underlying correlations between variables (Tufte 1983).

These visualizations provide a quick means of assessing data distributions, clusters, outliers, and correlations. Given a scatterplot, an analyst will visually assess the relationship between the variables being plotted by looking for trends in the plot. If the points tend to approximate a line running from the lower left to the upper right, this is often indicative of a positive correlation between variables. Likewise, if the points tend to approximate a line running from the upper left to the lower right, this can be indicative of a negative correlation. Such plots can also be enhanced by fitting a line to the data to help visualize such linear correlations.

However, (Cleveland 1993) notes though that putting a smooth curve through the data in the mind's eye is not a good method for assessing nonlinearity and can bias the analyst.

While scatterplots are good at assessing linear correlations, they can also be effective at implying higher-order correlations as well. Points in the plot may visually approximate other shapes, such as exponential or logarithmic curves. These early insights into the data can provide an excellent starting point for analysts and further generate new hypotheses about the relationships between variables within their data.

While scatterplots allow one to assess the bivariate relationship of data, data sources being analyzed today are typically multivariate. Common extensions to the scatterplot include encoding another variable to color the points on the plot or encoding a variable to the size of the points. However, even more commonly, a matrix of all possible variable combinations may be drawn such that each column of the matrix contains the same x-axis and each row contains the same y-axis for a given scatterplot. Such a matrix is called a *scatterplot matrix* and is useful for visualizing how a data set is distributed through multiple variables.

58.2.4 Parallel Coordinate Plots

While scatterplots can provide an overview of relationships between multivariate data dimensions, they are limited in terms of screen space and the fact that they only present two variable relationships at a time. Even by extending scatterplots to scatterplot matrices, the amount of screen space needed to represent the variables can still be inadequate for showing all possible relationships within a data set. In order to overcome such limitations for multidimensional data explorations, the *parallel coordinate plot* was introduced by (Inselberg 1985).

Parallel coordinate plots are comprised of a series of univariate data axes, with each axis representing some measurement within a data set. These axes are drawn parallel to each other as shown in Fig. 58.2. Each element of the data set is then represented as a line running through its corresponding axis value. As more variable attributes are added, more axes are added and more connections are made. In interpreting the parallel coordinate plots, the idea is that the user will recognize correlation values simply by the shape the plots make. If lines between two axes are close to parallel, they would have a correlation coefficient approximately equal to 1. If lines between two axes intersect one another in the middle of the graph, they tend to have a high negative correlation. If lines cross each other at different angles between two axes, the correlation coefficient between these two variables is approximately zero.

It is crucial to note that the ordering of the coordinate axes will play a major role in the analysis phase. In fact, effective dimensional ordering is a key component of many visualization techniques (e.g., star glyphs, pixel-oriented techniques), and a good ordering of the data can enhance the overall analysis process (Ankerst et al. 1998). Furthermore, the scaling of each axis and spacing also plays a major role in the analyst's ability to extract information from a parallel coordinate plot.

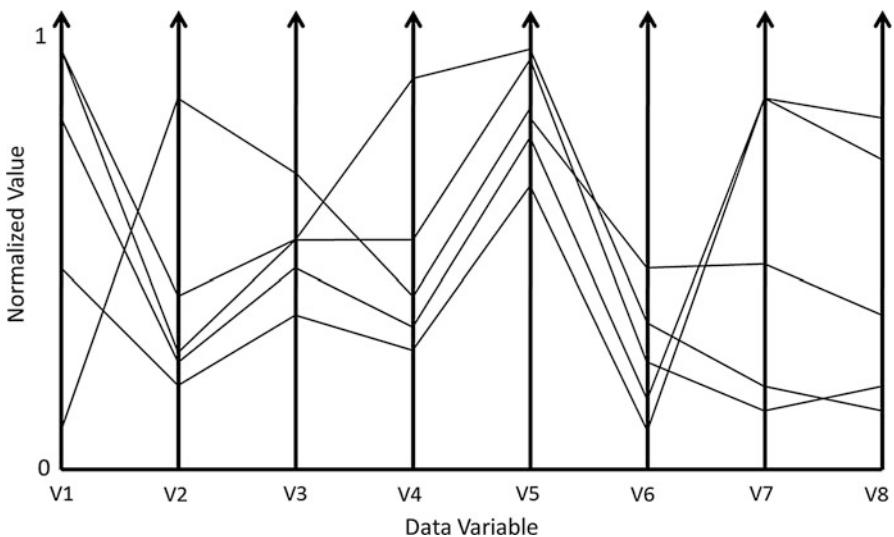


Fig. 58.2 A parallel coordinate plot representing six data attributes

Thus, while parallel coordinate plots are capable of representing more variables than a traditional scatterplot, more consideration needs to be taken when creating the visual representation.

58.3 Choropleth Maps

As discussed in the introduction, visualization is the process of creating interactive visual representations of data in order to generate knowledge about a particular process or phenomena found within the data. While any pictorial representation of the data may have the potential to generate knowledge, the question being asked of the data can directly influence what type of visualization could be most effective. In the previous sections, common statistical graphics were explored, each providing an overview of data distributions. However, in the previously explored visualizations, spatial relationships inherent within the data were ignored.

In geovisualization, often the first question being asked of the data is “compare location x to location y.” Such a question can easily be answered by plotting data elements on the map, and the user can visually explore such data for patterns. Data is often in the form of geographically reference latitude/longitude pairs, or aggregated into geographical areas on a map. Figure 58.3 illustrates various potential geographical visualizations of spatial data on a map. In Fig. 58.3a, locations of criminal offenses are plotted with respect to a centralized location denoted by the pin glyph and the semitransparent circle. In Fig. 58.3b, an aggregation of a pandemic influenza simulation is visualized, where each county is colored by the number of simulated ill patients. In Fig. 58.3c, the estimated probability

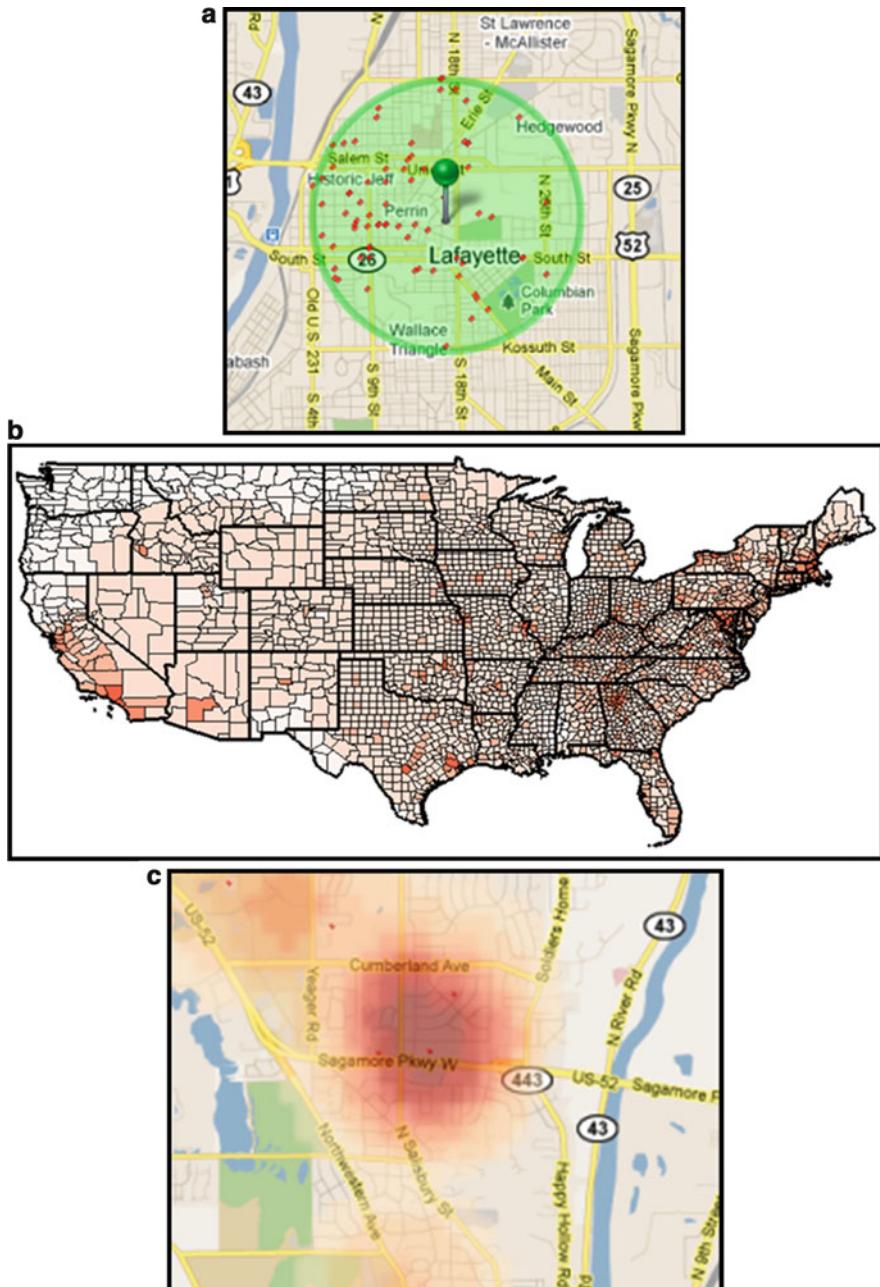


Fig. 58.3 Sample geographical data visualizations. (a) Plotting data as symbols/glyphs. (b) aggregating data by county in a choropleth map. (c) Abstracting and estimating data through density estimation

distribution of a criminal offense occurring is plotted, where the darker color represents a higher probability. Each of these representations is able to answer different questions about the data, and the underlying choice of an appropriate representation is crucial in constructing an effective geographic visualization.

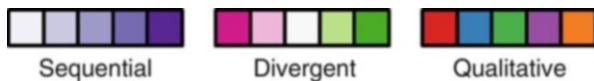
In order to construct an effective geographic visualization, designers must be aware that choices of glyph shape, size, and color can have a direct impact on how the information presented is perceived. Such choices are important not only in geographical representations, but in all visual representations of the data. In *The Semiology of Graphics* (Bertin 1967) and *The Grammar of Graphics* (Wilkinson 2005), the mapping of quantitative attributes (e.g., counts, rates, and other measures) to aesthetics (e.g., color, size, shape) is described in great detail. Of key importance, though, is how these mappings are perceived and interpreted by those using the visualization. The motivation behind this is that the graphics being displayed on the screen are the precise elements of which we are questioning. Thus, when choosing what visual aesthetic will be used, the designer first needs to consider what sort of data is going to be mapped to the aesthetic and how this may interact with other items being rendered. With regard to geovisualization, this interplay between cognition and perception for understanding how maps work is succinctly presented in *How Maps Work* (MacEachren 1995). In this work, MacEachren describes the ways in which the representational choices inherent in mapping interact with our information processing and knowledge construction to form insights into the data.

As such, proper design choices based on data attributes is of critical importance in creating a successful geovisualization. In order to illustrate the importance of such design choices, one can consider perhaps the most common type of thematic map, the choropleth map (see Fig. 58.3b). A choropleth map is a map in which areas are shaded or textured/patterned based on the measurement of a statistical variable in that region (e.g., income, population density). Such maps are based on data aggregated over previously defined regions (e.g., counties, states, countries) and are often used as a first-pass tool for exploring spatial statistics, for example, in exploring the data in Fig. 58.3b one may ask “where are the highest rates of illness?” However, questions such as that can be gleaned simply from a table of numbers, with the choropleth map, a user may now begin comparing rates spatially across regions. In order to effectively compare rates, the colors chosen to render the map need to be interpreted, and much research in both the cartographic and visualization domain has been done on what constitutes an effective color mapping for data values.

58.3.1 Color

The choice of the color scale is a complicated design choice that depends primarily on the data type, domain problem, and chosen visual representation. Typically, one will choose to constrain the color mapping to a univariate scale. While data is becoming increasingly multivariate, for example, the data set may have many

Fig. 58.4 Examples of univariate color maps for data visualization



different statistical measures within a county, many users often have difficulty interpreting bivariate or higher-dimensional color schemes. For a univariate color map, we can constrain ourselves to three types of color schemes (Harrouer and Brewer 2003): sequential, divergent, and qualitative (Fig. 58.4).

Each color scheme has its own strengths in regard to what type of data it can best represent. For data containing some sequential order (e.g., rates going from low to high), as the name suggests, a sequential color scheme is the most appropriate. In a sequential scheme, the color is mapped and ordered such that light colors (typically) represent lower values, and dark colors represent higher values. The divisions between each color band are chosen to be perceptually differentiable (Harrouer and Brewer 2003) and to also give the impression of increasing and decreasing values to the user.

The divergent color scheme functions in much the same way as the sequential color scheme. The key difference is the use of a comparison point or a zero point from which the data is being explored. In divergent scales, a key value of interest is mapped to the middle region of white, and values above and below this value are mapped along sequential color scales. Careful choices must be taken when choosing the high- and low-end representations for the scale. Often this is done with the concept of “cool” colors and “warm” colors as defined by Hardin and Maffi (1997), where red and yellow colors are considered warm and blues are considered cool.

The qualitative color scheme is unique in that the bands of this scheme are not perceptually ordered. Instead, the bands of color are chosen to be perceptually different and unorderable. This mapping would be used to map distinct data classes to a color, for example, counties that skew towards one political party would be one quantitative band, and those that skew to a different party could be another.

58.3.2 Class Intervals

While the choice of color scheme is critical in creating an appropriate rendering, the way the map is colored is based on a classification of the distribution of the variable being visualized. This classification is analogous to the previous discussion of creating histograms. Here, the number of colors being used is analogous to choosing the number of bins. First, the data over the entire geographical space can be analyzed, providing an overview of the data distribution. This distribution is then transformed into a histogram, where each bin of the resultant histogram will map to a particular color. Unfortunately, choosing the number of bins to represent the data is even more critical of a task when generating a choropleth map. If too many bins are chosen, analysts will be unable to distinguish the different color values mapped to different regions and may be distracted from seeing trends within their data. If too few bins are chosen, the overall trends will be over-smoothed, and patterns

can become hidden within this smoothing. Common methods for classification include quantile, equal interval, and standard deviation classifications; details and comparisons of methods can be found in (Monmonier 1972).

Once the number of colors is chosen (a good description of color maps and classifications is presented in Harrower and Brewer (2003)), then each geographical unit is colored based on its given statistical value. Along with the complexity of choosing categories for choropleth map colors, it is important to understand that size plays a dominant role in our perception. Geographic areas (e.g., counties, zip codes) vary in their shape and area (compare Alaska to Delaware), and a choropleth map is designed such that the map colors provide an equal representation to all areas within the map. Unfortunately, the different sizes of geographical areas can play perceptual tricks on the analyst, hiding changes in the data, or draw attention to unimportant areas of the map. Furthermore, when aggregating data, small areas (like major cities) may overwhelm the data of larger regions (like states). Such aggregation problems give rise to ecological fallacies as choropleth maps provide analysts with ample opportunity to make inferences on their data based on the aggregate region.

58.4 Exploratory Data Analysis

Until this point, the discussion has centered around generating statistical graphics and relatively simple choropleth maps. Each of these techniques has strengths and weaknesses; however, one can think of these techniques as individually being powerful tools for exploring data. The idea of using these tools as a means of investigating data was termed *exploratory data analysis* by Tukey (1977). Tukey compared the exploration of data to that of detective work in which a detective investigating a crime would need the tools necessary to analyze the crime scene (e.g., a finger printing kit) as well as an understanding of how crime works. He noted that data analysts need ways to look at their data, and whether these techniques are graphical, arithmetic, or in-between, the simpler they are made, the better they will be at conveying information to the user. Thus, we can think of the previously discussed visualization methods as a set of tools that can be used in creating geographical visualization systems.

Rather than trying to make the best map that can show all of our variables, it is perhaps best to expand from the notion of a single view. Instead, we should realize that data can be represented in a variety of ways, and given the highly multivariate nature of data being collected, a single map or statistical graphic may not be enough. Rather than trying to make one best view of the data, interactive graphics systems can provide multiple representations of the data. Furthermore, these representations can be programmatically *linked* or *coordinated*.

Figure 58.5 illustrates the concept of coordinating views. In this figure, an analyst is exploring criminal incident report data through a variety of displays. These coordinated multiple views (North and Shneiderman 2000) allow an analyst to create several displays of the data, typically involving statistical graphics and/or

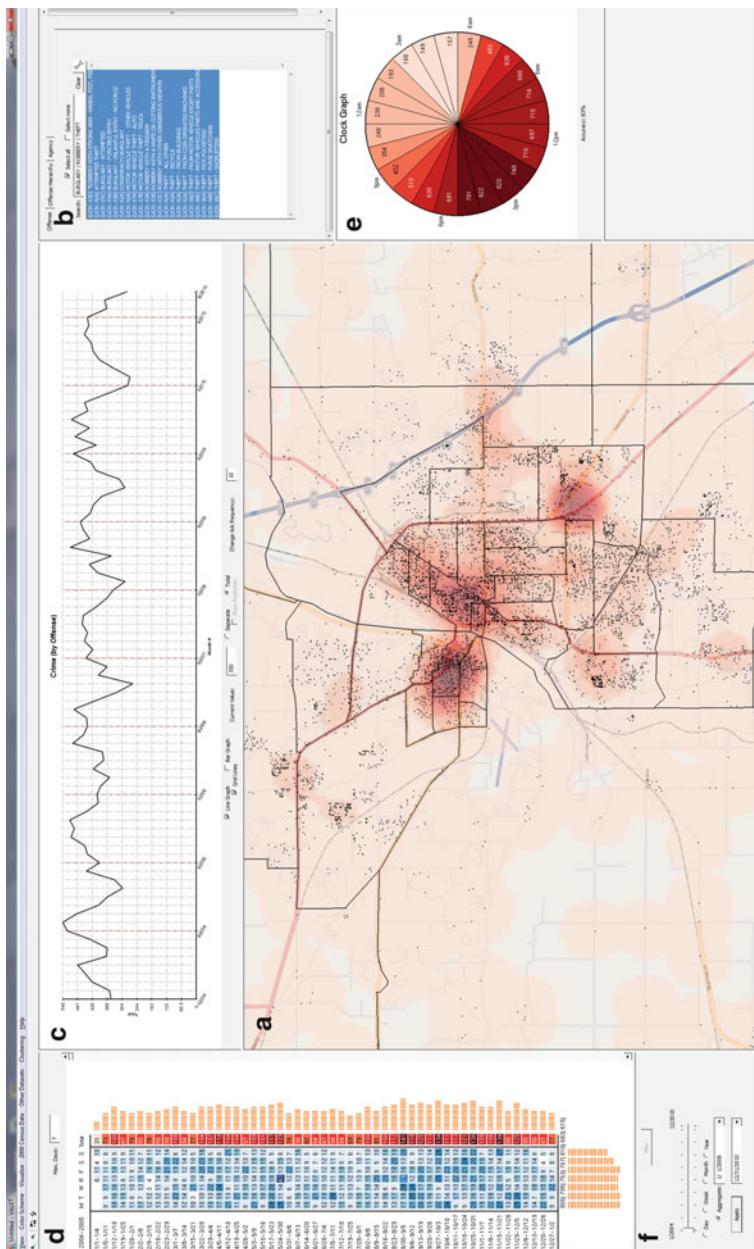


Fig. 58.5 The visual analytics law enforcement technology toolkit. Here, the user is exploring criminal incident reports using coordinated multiple views. The top-right window (b) shows an interactive menu that allows users to filter through multiple offenses/offense hierarchy/agencies. The top window (c) shows the time series and the left window (d) shows the calendar views of the selected incident report data. The right window (e) shows a clock view that provides an hourly temporal view of the data. The bottom-left window (f) shows the time slider with radio buttons that allow different temporal aggregation levels

geographical maps. Each display provides details on data entities, or aggregations of the data, and is linked through a variety of brushing and selection methods. It is through these interactions with the data that analysts are able to derive even more insight into their data. First, initial data views are typically constructed as a means of describing the data to the analyst, thus providing an overview of the data. An analyst then explores the data, generating and investigating hypotheses in a process that is often referred to as the *Visual Information Seeking Mantra* coined by Shneiderman (1996): “Overview first, zoom and filter, then details on demand.”

Perhaps the most basic interaction techniques are those of *scrolling and panning*. Given a large data set, it is possible that some data will be rendered outside of the screen space. In order to explore this data, the user will need to move the display by scrolling or panning in order to bring other data elements into view. These basic interaction techniques spatially separate the current focus from the larger information space by providing a sliding virtual window. However, scrolling/panning interactions introduce a cognitive burden upon the user where the analyst must now keep track of the context of their data.

Similarly, *zooming* techniques temporally separate focus regions from the information context, inducing a virtual hierarchical ordering at various zoom levels. Zooming also places a significant cognitive burden on the user to assimilate focus into the overall information space as the various zoom levels may not provide any explicit contextual cues. This causes problems such as “desert fog” (Jul and Furnas 1998) in which a view of the data information provides no details on which to base further navigations within the data set.

In order to reduce the cognitive load, *overview + detail* and *focus + context* techniques were introduced. Overview + detail techniques attempt to reduce the cognitive burden associated with exploratory navigation (i.e., panning and zooming) by providing simultaneous synchronized views of overview and detail of the information space. An example would be the thumbnail views provided in Adobe Reader and Microsoft Powerpoint. Focus + context techniques embed the focus region within the larger information space using a transition function in order to overcome issues with cognitive reorientation between overview and detail. Fisheye lenses (Furnas 1986), implemented using a distortion-based transition function, are a commonly used design choice for focus + context techniques. Evaluations of focus + context techniques, however, suggest that they are not always beneficial (Gutwin and Skopik 2003). Object targeting becomes difficult due to “hunting effects” of the fisheye that occurs as a result of magnification. Fisheye distortions have also been found to interfere with a user’s spatial comprehension as well as with other tasks involving location recall and visual scanning.

While such interactions are useful for exploring details within the data, perhaps the most critical interaction is that of *brushing*. Given a statistical graphic, the user can interactively select (brush) data elements, for example, histogram bars, points on a scatterplot, and lines on a parallel coordinate plot. If several different views are linked to the same data source, this brushing will highlight these elements across all views. (Monmonier 1989) further expanded this notion of integrating brushing (particularly scatterplot brushing) with maps, calling this a geographic brush.

By selecting areas of the map, points in the scatterplot would be highlighted, and vice versa. Examples of brushing in coordinated multiple views are shown in Fig. 58.6.

By employing interaction techniques such as brushing, linking, and drill-down operations, users can now explore areas on the graphic and retrieve the exact data values. Furthermore the use of interactive techniques for exploring data graphics has several advantages over traditional static graphics. The first is that such interactivity will allow for greater precision. For example, when analyzing a static map, colors represent a range of value; the addition of interactivity allows one to query the exact value of a region. Second, this interaction not only adds precision but also provides users with a quicker means of data retrieval. Attention is no longer split between looking at the graphic and looking at the legend or scale; instead, users can click regions to retrieve data values. Such techniques all fall under the umbrella of exploratory data analysis, and statisticians within geography have further expanded on these ideas, terming them *exploratory spatial data analysis*. A review of exploratory spatial data analysis can be found in Anselin (1998) and Andrienko and Andrienko (2006).

Exploratory spatial data analysis utilizes the same tools and interaction techniques listed above; however, the focus is primarily on spatial or spatiotemporal data. As such, exploratory spatial data analysis systems link tools such as scatterplots, histograms, and others to interactive maps. The links between these statistical graphics are through brushing and highlighting, where users can interactively select a portion of the data in one view and see these elements highlighted in other views. In this way, users can begin developing and exploring hypotheses about their data. Such tools are then linked to analytic algorithms that can provide deeper insight into data correlations and statistics. These methods focus explicitly on the spatial aspects of the data (spatial dependence, association, and heterogeneity). The goal is to discover spatial patterns and relationships within the data by combining a variety of exploratory data analysis techniques with spatial analysis algorithms and geographic information system tools. Currently, a variety of exploratory spatial data analysis tools are available online, and details of these systems can be found in *Handbook of Applied Spatial Analysis* (Fischer and Getis 2010).

58.5 Exploring Time

The discussion up to this point has focused on visualizations for summarizing data and exploring spatial components of the data. However, we do not want to constrain geovisualization to only spatial data, as often data being collected geographically contains information about rates, movements, and changes over space and time. Instead, we want to create systems and visualizations that are able to answer questions not only about where but also when, and by combining representations that can answer these sorts of questions, we can begin generating insight into how or why something is occurring. In fact, many questions asked about temporal data have to do with the change throughout time. These patterns are formed by the combination of four characteristics (Few 2009): the magnitude of the change, the

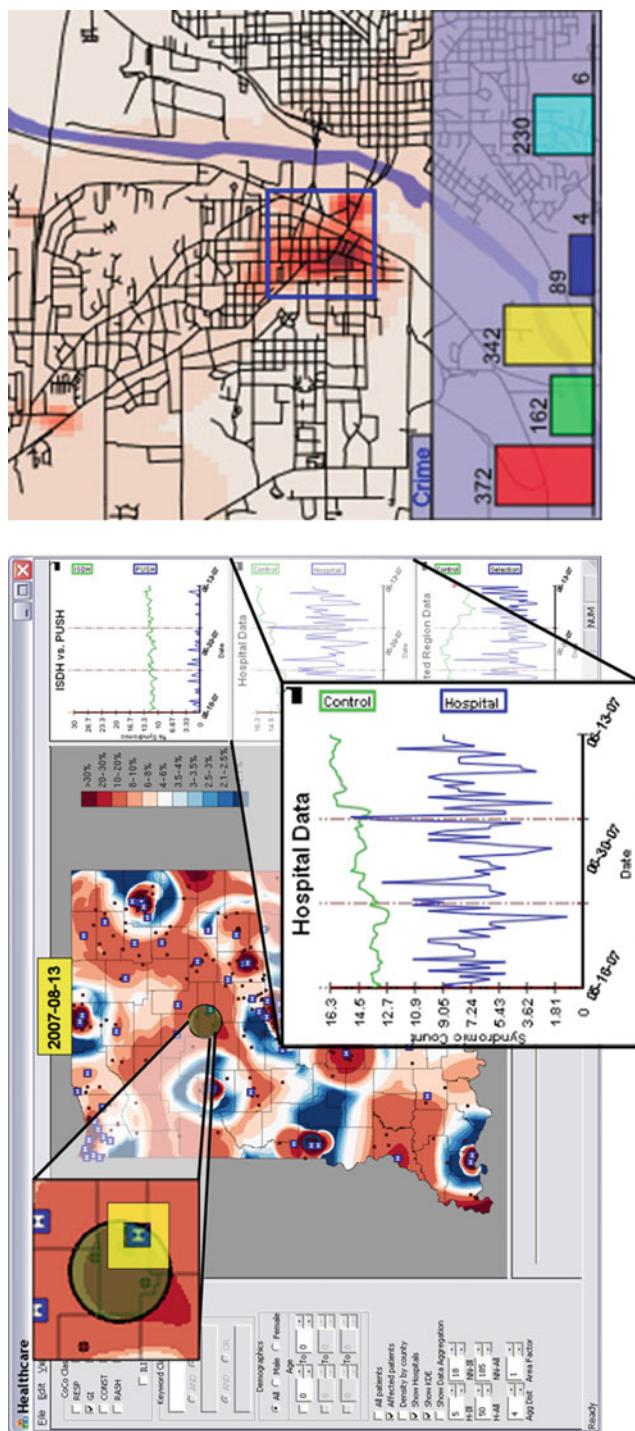
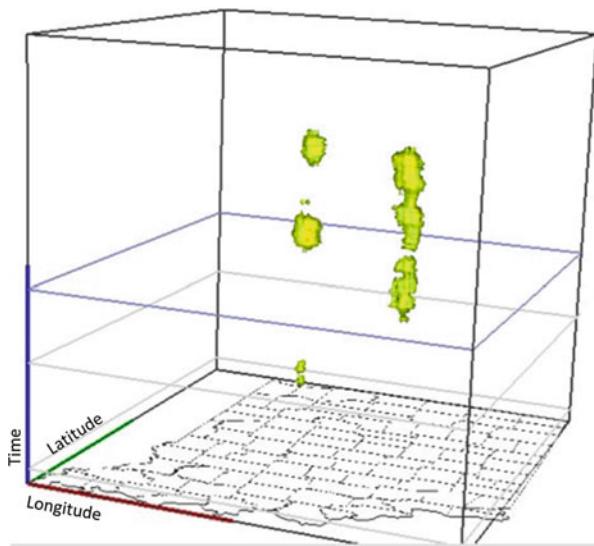


Fig. 58.6 Brushing in coordinated multiple views. (*Left*) exploring hospital patient records using linked time series plots. (*Right*) exploring criminal incident reports using histograms

Fig. 58.7 An example space-time cube. Disease rates are mapped over space and time to explore how spread vectors may occur and what clusters are potentially of interest



shape of the change, the velocity of the change, and the direction of the change. A recent review of temporal visualization methods can be found in Aigner et al. (2008); however, this section will focus on two techniques primarily used for spatiotemporal data (as opposed to strictly temporal data).

58.5.1 Animation

One of the most common ways of displaying spatiotemporal data is through the use of animation. Since time moves in a linear fashion, one can animate graphics to show the movement of trends over time. Unfortunately, the animation of choropleth maps brings with it a series of new challenges. In previous sections, a discussion on class interval selection was provided, and the choice of class interval looks at data for only one given time interval (or aggregate thereof). However, when the statistics are allowed to change temporally, the choice of class intervals becomes increasingly challenging. Now, the class choice must work not only across one map, but across multiple maps, and the choice of class interval can potentially emphasize relatively small fluctuations due to temporally global choices in class selection. While such issues complicate the creation of an animated choropleth map, many systems allow for either looped playback animation or user-controlled exploration through an interactive time slider.

58.5.2 Space-Time Cube

While animation provides an obvious way to display spatiotemporal data, it also introduces cognitive burdens onto the user similar as the user now must retain

information of the last state of the data visualization and compare it to the current state. One way of removing such a burden would be to display both space and time concurrently in a visualization. At the end of the 1960s, Hägerstrand (1970) introduced a geographical technique called the *space-time cube*. This technique utilized a three-dimensional diagram (or space-time cube) to visualize spatiotemporal data. The space-time cube consists of a two-dimensional geographical space and a third dimension of time. In this way, spatiotemporal data can be visualized showing movement patterns and trends in a single graphical representation. Figure 58.7 is a space-time cube visualization of disease rates, showing clusters over space and time. Like animation, space-time cubes introduce their own cognitive burdens. When rendering in three dimensions, data occlusion can occur, and rotation, panning, and zooming becomes necessary, thus creating a larger cognitive load.

58.6 Conclusion

Currently, data is being generated and collected at unprecedented rates, particularly georeferenced data. Cell phone locations, georeferenced tweets, street cameras, and others all provide analysts with new streams of data containing locations about incidents that may or may not be of interest. Furthermore, with the development and growing popularity of *geobrowsers*, more and more individuals can now readily map their own georeferenced data. For example, Google Maps provides an API (application programming interface) in which users can quickly plot and browse their georeferenced data in a sophisticated and intuitive user interface. Such tools provide opportunities to explore data from bike traffic routes to crime patterns to individualized restaurant preferences.

In exploring current research in data visualization, it becomes clear that geographical visualization is playing a larger and larger role as a central piece in data analysis and exploration. Geovisualization is not purely about creating maps of data. It is about generating tools and techniques for visually representing spatial and spatiotemporal data and facilitating the exploration of this data for hypothesis generation and exploration. The next step in geovisualization is to expand from hypothesis generation and exploration and begin incorporating tools for data modeling and hypothesis testing. Currently analysis tools (e.g., Anselin et al. 2006) are incorporating both interactive graphics and advanced statistical analysis algorithms for data exploration and analysis.

Given the ability of the human visual system to recognize patterns within data sets, it is imperative to allow analysts to explore and interact with their data. In the problems given here, data analysis was presented as more of a searching and hypothesis generation problem; however, it is often the case that data sets have been generated with specific domain questions in mind. By tailoring our analysis, visualizations, and interactions to these domains, it is possible for an analyst to gain more insight into their data than would have been possible with static graphics or traditional tools. Furthermore, this chapter has discussed only the most basic components of visual analysis. Hypothesis testing algorithms and data mining tools exist that allow

analysts to readily sift through their data; these techniques are being coupled with novel graphical displays in an attempt to generate large amounts of insight into data. The details presented here form only the beginnings of geographical visualization.

Current research explores geometrical modeling of cartographic techniques such as flow maps. Notations, streamlines, and other visual cues have been added to space-time cube visualizations as a means of enhancing information. Three-dimensional density maps can be easily generated and explored using interactive techniques, and combinations of various textures, glyphs, symbols, and colors are being explored as a way to represent larger amounts of data to a user in a single display. These visuals are used not only in the analysis and exploration process but also as a means of explaining data trends and narratives to others. With geographical visualization, the general populace understands a map, and by linking their data to these sorts of displays, they become invested in the analysis. However, care must be taken when choosing our graphical representations. It is easy to lie with statistics, and as shown here, statistics is one of the foundations of data visualization. As such, we should strive for creating reliable and correct representations of data, while providing analysts with interactive means of exploring and ultimately analyzing their data.

References

- Aigner W, Miksch S, Muller W, Schumann H, Tominski C (2008) Visual methods for analyzing time-oriented data. *IEEE Trans Vis Comput Graph* 14(1):47–60
- Andrienko N, Andrienko G (2006) Exploratory analysis of spatial and temporal data: a systematic approach. Springer, Berlin
- Ankerst M, Berchtold B, Keim DA (1998) Similarity clustering of dimensions for an enhanced visualization of multidimensional data. In: IEEE symposium on information visualization, Phoenix, pp 52–62
- Anselin L (1998) Exploratory spatial data analysis in a geocomputational environment. In: Longley PA, Brooks SM, McDonnell R, MacMillan B (eds) Geocomputation: a primer. Wiley, Chichester, pp 77–94
- Anselin L, Syabri I, Kho Y (2006) GeoDa: an introduction to spatial data analysis. *Geogr Anal* 38(1):5–22
- Bertin J (1967) The semiology of graphics. ESRI Press, Redlands
- Cleveland WS (1985) The elements of graphing data. Wadsworth, Monterey
- Cleveland WS (1993) Visualizing data. Hobart Press, Summit
- Few S (2009) Now you see it: simple visualization techniques for quantitative analysis. Analytics Press, Oakland
- Fischer MM, Getis A (eds) (2010) Handbook of applied spatial analysis. Software tools, methods and applications. Springer, Berlin/Heidelberg/New York
- Furnas G (1986) Generalized fisheye views. In: Proceedings of ACM CHI, Boston, pp 16–23
- Gutwin C, Skopik A (2003) Fisheyes are good for large steering tasks. In: Proceedings of the SIGCHI conference on human factors in computing systems, Ft. Lauderdale, pp 201–208
- Hägerstrand T (1970) What about people in regional science? *Pap Reg Sci* 24(1):6–21
- Hardin C, Maffi L (1997) Color categories in thought and language. Cambridge University Press, Cambridge
- Harrover MA, Brewer CA (2003) ColorBrewer.org: an online tool for selecting color schemes for maps. *Cartogr J* 40(1):27–37
- Inselberg A (1985) The plane with parallel coordinates. *Vis Comput* 1(4):69–91

- Jul S, Furnas GW (1998) Critical zones in desert fog: aids to multiscale navigation. In: Proceedings of the 11th annual ACM symposium on user interface software and technology, San Francisco, pp 97–106
- MacEachren AM (1994) Visualization in modern cartography: setting the agenda. In: MacEachren AM, Taylor DR (eds) *Visualization in modern cartography*. Pergamon, Oxford, pp 1–12
- MacEachren AM (1995) How maps work. Guilford, New York
- Monmonier MS (1972) Contiguity-biased class-interval selection: a method for simplifying patterns on statistical maps. *Geogr Rev* 62(2):203–228
- Monmonier M (1989) Geographic brushing: enhancing exploratory analysis of the scatterplot matrix. *Geogr Anal* 21(1):81–84
- North C, Shneiderman B (2000) Snap-together visualization: evaluating coordination usage and construction. *Int J Hum Comput Stud* 51:715–739
- Pearson K (1895) Contributions to the mathematical theory of evolution. II. Skew variation in homogenous material. *Philos Trans R Soc A Math Phys Eng Sci* 186:326–343
- Shneiderman B (1996) The eyes have it: a task by data type taxonomy for information visualizations. In: Proceedings of the IEEE symposium on visual languages, Las Alamos, pp 336–343
- Sturges HA (1926) The choice of a class interval. *J Am Stat Assoc* 21(153):65–66
- Tufte ER (1983) *The visual display of quantitative information*. Graphics Press, Cheshire
- Tukey JW (1977) *Exploratory data analysis*. Addison-Wesley, Reading
- Wilkinson L (2005) *The grammar of graphics*. Springer, New York

David Manley

Contents

59.1	Introduction	1158
59.2	MAUP Definitions	1159
59.2.1	The Scale Effect	1159
59.2.2	The Zonation Effect	1160
59.3	Approaches to Understanding	1162
59.3.1	From Univariate Statistics to Spatial Models	1162
59.3.2	The Importance of Spatial Autocorrelation	1165
59.3.3	Exploring the MAUP Through Zone Design	1167
59.4	Conclusion	1169
	References	1170

Abstract

The modifiable areal unit problem (MAUP) is a serious analytical issue for analysts using spatial data. The MAUP manifests itself through the instability of a wide range of statistical results derived from analysis on spatially organized data. When spatial data are aggregated, the results are conditional on the spatial scale at which they are conducted, and the configuration of the areal units that are employed to represent the data. Such uncertainty means that the results of spatial data where the MAUP has not been considered explicitly should be treated with caution. Although solutions have been proposed, none have been applicable in more than a couple of specific cases. As such, it is likely that the MAUP will never be truly solved. This chapter charts the two related aspects of the MAUP, the scale and zonation effects, and details the role of spatial autocorrelation in understanding the processes in the data that lead to the

D. Manley

School of Geographical Sciences, University of Bristol, Bristol, UK

e-mail: d.manley@bristol.ac.uk

statistical nonstationarity. The role of zone design as a tool to enhance analysis is explored and reference made to analyses that have adopted explicit spatial frameworks.

59.1 Introduction

A serious problem for analysts of spatial data is that while the phenomena they are investigating may be continuous, the data available frequently are not, and the areal units used to present the continuous data are arbitrary compromises designed to suit a wide range of uses rather than spatial equivalents of the day, month, or year. As a consequence, statistical analysis of individual data that has been aggregated into areal units is susceptible to nonstationarity across a wide range of measures. This problem is known as the modifiable areal unit problem (MAUP), and it has vexed users of aggregate data for many decades. Countless investigations have demonstrated that it is unlikely that an analytical solution to the MAUP will be identified, and those solutions that have been proposed frequently suffer from substantial flaws. Indeed, as yet, we have neither a full and detailed understanding of the problem nor the underlying causes. It is unlikely that an analytical solution to the MAUP will ever be realized due to the wide range of possibilities that arise when the partitioning of continuous space is implemented as well as the wide range of analytical tasks that aggregated data are required to perform (for comprehensive overviews of the MAUP, see Openshaw 1984; Wong 2009). Instead, the MAUP needs to be accounted for clearly in the research hypothesis that precedes analysis. In the twenty-first century, spatial data are an increasingly important factor in everyday life. Almost all nations in the developed world collect and publish data using administrative boundary systems – areal units. In the United Kingdom, the decennial population census is published using small, low-level areal units. Small area geographies, for a comprehensive range of area characteristics such as are available for the British Census, are valuable as the hidden aspects of the problem are less likely to occur, other things being equal, at fine levels of granularity than coarse ones. It is also worth noting that the small areal units of the British Census were designed explicitly drawing on the principles of the MAUP, promoting, amongst other things, internal homogeneity across a range of important indicators such as housing tenure. The problem of the MAUP is magnified by the temporary nature of the areal units and the frequent revisions that are made to the coverages to reflect changes in population data.

Despite the prevalence of the MAUP in spatial data, it is an issue that is all too frequently ignored or neglected in geographical analysis. A search in Google Scholar on the term “modifiable areal unit problem” reveals only 4,160 publications, a low number when you consider the number of papers that deal with aggregated data in their analysis (around 400,000). The lack of attention paid to the MAUP has, perhaps, two underlying causes. Firstly, the readily available nature of many areal unit systems means that the majority of research using aggregate data adopts areal boundaries that are generated *a priori* and an engagement with the creation of areal units is not required. Secondly, the results of many quantitative studies that employ aggregate data of one

sort or another rely on the implicit assumption that the MAUP isn't a significant problem in order to present valid results. To acknowledge the MAUP, even informally, would be to question the validity of the analysis conducted and conclusions reached. Openshaw's conclusion from almost three decades ago remains as pertinent today as it was when he wrote it: "this is hardly a satisfactory basis for the application and further development of spatial analysis techniques in geography" (1984, p. 5).

This chapter explores the problem of the MAUP in the context of spatial data analysis, outlining the two major aspects of the problem, the scale effect and the zonation effect. Definitions are provided for both these aspects, and examples are drawn from the literature to illustrate the problems. Following these two sections, an overview of the evidence relating to the MAUP is provided.

59.2 MAUP Definitions

There are two aspects to the MAUP known as the scale effect and the zonation effect (also called the aggregation effect in some literature (for instance Openshaw 1977), but since the process of aggregation is involved in both scale and zonation decisions, an important distinction is made here, and the term zonation effect employed). This section outlines the two aspects with reference to relevant examples and provides the context for a discussion around empirical results in the following sections.

59.2.1 The Scale Effect

The scale effect arises because of the nested hierarchies within which human society is arranged and is expressed through the task of choosing the most appropriate scale for analysis (Arbia 1989) (Fig. 59.1). It is rarely that clear at which spatial scale an analysis should proceed, and frequently, there are multiple spatial scales at which an analysis could theoretically be conducted. Drawing on the United Kingdom Census as an example, output areas (OAs, typically 140 individuals) form the basic spatial units and can be aggregated into higher-level spatial units, such as wards (usually a couple of 1,000 individuals) and districts (many 100,000s of individuals).

The "classic" example of the scale effect was published by Gehlke and Biehl (1934) and used three different datasets including random coin tosses, census data, and experimental groups of rural counties drawn from the United States (see also Yule and Kendal 1949). They demonstrated that coefficients from correlation analyses between, for instance, census data reporting juvenile delinquency and monthly house rentals tended to increase as the number of areal units representing the data decreased. Table 59.1 reproduces the results of their correlation analysis. While the census data may be susceptible to structures within the data that cannot be observed, which in turn cause the instability of the statistical results, the coin toss data demonstrated that correlation coefficients changed even when the underlying data were generated randomly, and each data unit was independent of all others. From their analysis, Gehlke and Biehl concluded by questioning whether or not

Fig. 59.1 The scale problem: The three different scales could represent (a) output areas, (b) wards, and (c) districts

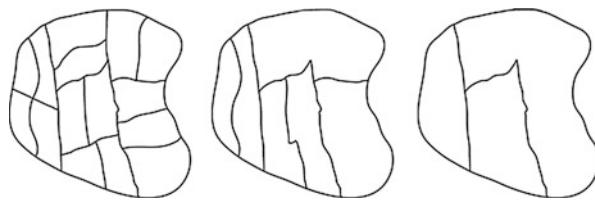


Table 59.1 Correlation coefficients under aggregation using juvenile delinquency and monthly rentals (from Gehlke and Biehl 1934, p. 169)

Number of areal units	Correlation coefficient (r)
252	-0.502
200	-0.569
175	-0.580
150	-0.606
125	-0.662
100	-0.667
50	-0.685
25	-0.763

“a geographical area is an entity possessing traits, or merely one characteristic of a trait itself” (p. 170). In essence, they urge caution in the treatment of data from areal units and “that variations in the size of the correlation coefficient seemed conditioned on the changes in the size of the unit used” (op.cit.).

Exploring the scale effect, Kirby and Taylor (1976) use data on referendum voting patterns to illustrate the potential pitfalls and identify pockets of the population who vote differently to the overall outcome for an area. The implication of this finding being that if analysis is conducted at difference scales it is possible to produce different area results from a single pattern of individuals voting. Kirby and Taylor also discuss the dilemma of choice of scale: at a scale that is too small, then it is not possible to compare data sources from different (modifiable) unit systems. However, with the scale too large, then much of the more local-level detail within an analysis is lost through the aggregation process. The scale effect has, therefore, a number of different elements, including the enhancing or smoothing of spatial processes, akin to the statistical smoothing of data to remove noise. The nontrivial nature of the scale effect was emphasized by Openshaw (1984), noting that even a relatively small set of zones can produce a sizable range of combinations: for instance, combining 1,000 zones into a new system of just 20 groups produces 10^{1260} unique combinations!

59.2.2 The Zonation Effect

Once the scale of the zonal system has been determined, then we can consider how the space is to be divided up – the zonation effect. The zonation effect occurs where there are

Fig. 59.2 The zonation problem. Each of these diagrams demonstrates a division of a sample space into five distinct areal units, yet each could potentially yield different results

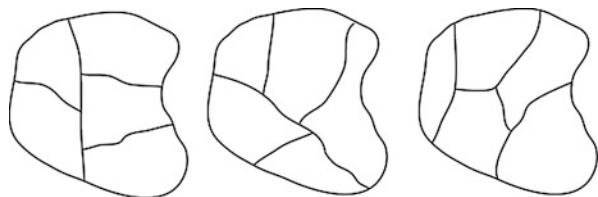


Table 59.2 Correlation coefficients from Openshaw and Taylor (1979, p. 129) showing zonation effect (adapted)

Number of areal units	Correlation coefficient (r)
Six republican-proposed	0.482
Six democratic-proposed	0.627
Six congressional districts	0.265
Six urban/rural regional types	0.862
Six functional regions	0.713

“any variations in results due to alternative units of analysis where . . . the number of units, is constant” (Openshaw and Taylor 1979). There are potentially an infinite number of different ways in which a continuous space can be subdivided into discrete areal units. A diagrammatic interpretation of the zonation problem is presented in Fig. 59.2.

For Openshaw (1984) the zonation effect was by far the greater of the two aspects of the MAUP, as there is considerably more freedom choosing the delineation of boundaries than in choosing the number of zones required. The consequence of this is that “the process of zonation becomes susceptible to the whims of those involved in the overall aggregation process” (Openshaw and Taylor 1981, p. 61). While this position may be extreme, it makes the point that there are serious problems with the arbitrary nature of the many areal units.

Openshaw and Taylor (1979, 1981) conducted one of the largest investigations into the MAUP. Replicating the earlier work of Gehlke and Biehl, they used correlation analysis to assess the instability of statistical analysis as a consequence of the MAUP. In the first instance, they correlated the proportion of republican voters against the percentage of the population above 65, using the 1970 US Census. To assess the impact of the zonation effect, Openshaw and Taylor produced correlation coefficients for multiple arrangements of counties in the state of Iowa. They set the scale constant each time aggregating the base units into six counties. Table 59.2 reports the results of their analysis.

Openshaw and Taylor (1979) demonstrated that it was possible to obtain highly changeable correlation coefficients for a single set of data. They went further than this in the article by attempting to describe the universe of correlation coefficients that were possible to achieve using the different scales of zonation. For many of the scales, they claim that the theoretical range of coefficient was from -0.999 to 0.999. However, this was rarely the case for many of the zonation systems that they devised.

For instance, using 72 zones the minimum found was -0.579 , and the maximum was 0.927 . This demonstrates the impact of the zonation effect as differing boundary choices change the correlation coefficient values.

59.3 Approaches to Understanding

There is a vast body of research that has sought to gain greater understanding about the MAUP and how it can impact the results of statistical analysis. This section reviews work that has sought to unpick how the MAUP can lead to different results in statistical analysis. Starting with the simple examples of uni- and bivariate analysis, evidence is provided that shows the potential severity of the MAUP. This is built on by introducing research that has examined the impact of the MAUP in models that are more explicitly spatial in configuration. Attention is then paid to the role of spatial autocorrelation and spatial cross-correlation, two of the fundamental processes that lie behind incidence of the MAUP. Finally, attention is paid to the process of zonation, or zone design, and the research that has been undertaken to explore the MAUP from the perspective of the aggregation process.

59.3.1 From Univariate Statistics to Spatial Models

There are many examples of investigations into univariate and bivariate parameter instability as a consequence of the MAUP. In a recent article that clearly demonstrates the importance of preserving the availability of small area estimates for understanding societal processes, Flowerdew (2011) took 2001 Census data for England and presented an investigation on the severity of the MAUP. While there are many studies that demonstrate that it is possible to obtain different statistical results for different spatial scales and configurations, there are fewer studies that then provide statistical evidence that these differences are significant. Developing this theme and 18 common variables, Flowerdew demonstrates that even just using the three standard spatial scales that the data are released at leads to results with significant statistical differences. Flowerdew uses the Fisher transformation to standardize the correlation coefficients and concludes that after standardization the MAUP effect leads to different results in around 60 % of the cases. In general, under increasing scale aggregation, the increase in correlation coefficient is a consequence of the data smoothing properties associated with the aggregation process. As such, the variation between variables tends to decrease as aggregation increases – the heterogeneity *between* units will fall as greater number of the population are combined into single entities and the heterogeneity *within* units increases.

There are fewer examples of investigations into the MAUP in multivariate analysis. However, Fotheringham and Wong (1991) did tackle this problem using American Census data and demonstrated the problem with a regression model that related mean family income for multiple unit configurations at various spatial scales.

As with the work presented above, Fotheringham and Wong demonstrate that different spatial scales lead to systematic variability in the outcome of the regression analysis so that for some parameters (percentage elderly and percentage blue collar workers) the relationship to mean household income becomes more negative. Conversely, other parameters (percentage of home owners and the percentage of black residents) become systematically less negative as aggregation increases. Again, the importance of the MAUP is demonstrated by investigating whether the differences in the parameters obtained at the different scales and zonal configurations are statistically significant. Fotheringham and Wong use distribution of the parameter estimates and judge significance using a standard difference means test with 1.5 standard deviations (the 95% confidence level) as the key cutoff point. As they conclude, there are “many places[...] where the parameter estimates are significantly different” (1991, p. 1035).

Although the models above consider the MAUP, none of them explicitly incorporate the spatial structure of the data. Moving beyond these aspatial models requires a more complex modeling strategy and the explicit adoption of a spatial framework. An example of a spatial investigation into the MAUP has been conducted by Baumann and colleagues (1983). In their work, they investigate what they term as the scale hypothesis and the aggregation hypothesis (the scale effect and the zonation effect in other words) with respect to the supply of labor in multiregional labor markets. Adopting a standard MAUP approach, they suggest that the way in which the model of labor supply is measured through participation rates and commuting flows may be affected by the scale at which an analysis is conducted and the regions through which the multiple labor markets are realized. In their findings, Baumann and colleagues present a number of interesting outcomes: firstly, in terms of determining labor participation (the number of males and females in employment), the effects of scale are relatively small. Thus, there is little variation in the result as the spatial scale of the analysis is altered. However, in a model representing commuting patterns, the scale effects are much larger, a finding which intuitively makes sense as commuting is only realized in the framework when zone boundaries are crossed. Increasing the scale will, all other things being equal, reduce the number of boundaries and so the level of commuting. In surmising their findings, Baumann and colleagues highlight that the spatial framework that is adopted for an analysis is crucial, and it is “by no means admissible to ignore possible effects of the choice of a spatial framework in spatial model building” (p. 67). Finally, they suggest that when seeking out the most appropriate spatial framework, a range of criteria including model R^2 , t-values, and a priori signs should be considered. This might lead the analyst to conclude, therefore, that the most appropriate spatial framework would be one that leads to the greatest level of explanation in the final model the best model performance overall. Within an econometric framework, this is an entirely reasonable assertion.

A major area of interest where the spatial organization of individual units within and between areas is segregation (see also Poulsen et al. 2011). It is a highly spatial phenomenon, and there are many examples within the literature where spatial statistics have been used to attempt to understand the role that the definition of

the areal units and the scale of analysis can have on the resulting measures. Wong's investigation into segregation indices and the MAUP demonstrated that, in general, as the spatial resolution (scale) increases, the greater the degree of segregation identified (Wong 2003). As discussed above, the scale process is akin to data smoothing, so that sharp inconsistencies between smaller units are removed. Thus, as the areal units become smaller, the potential level of homogeneity within the areal unit will increase because there are fewer individual data points represented within each unit (up until the level of the single individual atomistic unit beyond which it is no longer realistic to decompose and represent a perfectly homogenous social unit). Using multiple scales of aggregation, Wong demonstrates that different scales produce different results for the dissimilarity index, D (see Duncan and Duncan 1955). To understand the impact of the MAUP scale effect, Wong proposes that the index can be decomposed into regional and local effects and that the local-level measure demonstrates the deviation of each unit from the global regional D value. The range of values achieved can give insight into how much each local unit influences the overall segregation pattern. High values record areas that deviate substantially from the global regional value, while lower values demonstrate congruence. Of course, one influence that Wong does not attempt to cover is the effect of zonation differences. It is clear however that with a small extension it would be possible to use Wong's methodology to effectively assess the impact of altering the boundaries on the resulting segregation outcomes. A second example using the diversity index, H, is used to highlight that with modification, it is possible extend the decomposition process to other segregation measures.

Two further examples of the MAUP impacting on the results of spatial statistical analysis are provided by the health literature, where research into the MAUP has been particularly active. The first study investigated the effects of the Dounreay Nuclear Power Plant in relation to instances of childhood leukemia as part of a public inquiry into an application to introduce reprocessing facilities (Heasman et al. 1984). In close proximity to the Dounreay plant were apparently high incidences of childhood leukemia. To investigate whether or not these represented significant clusters of leukemia in children, the Scottish Health Service analyzed data recording all incidences of cancer between 1968 and 1986. The initial results of the analysis reported that there was a significant excess of cases in the Dounreay area. However, at the subsequent public inquiry, a number of methodological weaknesses were identified, amongst which was the issue of boundary definition, the MAUP. Wilkie (1986) provided details of the methodological problems which included the potential gerrymandering (manipulation) of the time period studied and radial distances used to detect the cancer clusters. Creating tight boundaries around cancer points would have the effect of forcing the mortality rates upward, creating artificially high results because of the smaller population bases. Similarly, looking at a different time period, either by cutting the time series data into different lengths or curtailing the investigation at an earlier time point, would have the effect of altering the outcomes observed. Further problems arise from the presence of edge effects (cases appearing near the edge of the study space) and irregularly

shaped areal units used for the aggregation. Finally, the use of areal units as a means to imprecisely locate individual incidence data introduced small errors which cumulatively could result in the erroneous generation of clusters where there were not any, or vice versa. In conclusion, the findings of the Dounreay analysis were difficult to evaluate robustly as the choice of radii and time periods for their study area “are arbitrary” (p. 266). Any clusters of cases in one area and time period could be eliminated simply through an alternative choice of radii or time periods. The second of the health examples is provided by Odoi and colleagues (2003). They were investigating the impact of the MAUP on the spatial distribution of human giardiasis (a parasitic infection causing diarrhea) in Canada. The study sets out to explicitly examine the impact of alternative spatial scales on the identification of infection clusters and whether the most appropriate statistical framework for assessing the clustering was using global or local statistics. Their analysis demonstrated that using a fine spatial scale with relatively small units enabled the detection of clusters that were hidden at the higher spatial scale. They also identified that local statistical measures provided more clustering detail than the global measures and as such were more appropriate for the exploratory analysis of patterns in spatial data.

59.3.2 The Importance of Spatial Autocorrelation

Tobler’s First Law of Geography states that all things are related, but near objects are more related than distant objects (Tobler 1970). More formally, the degree of similarity is known as spatial autocorrelation, a concept developed by Michael Dacey in the 1950s at the University of Washington (see Getis 2010 for a comprehensive review). Cliff and Ord (1981) make the link between spatial autocorrelation and the MAUP more explicit, and note that the size of the cells in the areal unit system is important in determining the strength of the spatial autocorrelation. All other things being equal, larger areal units will have lower levels of autocorrelation than smaller ones. In other words, at different spatial scale, different patterns and degrees of spatial autocorrelation will be present and will impact on the structure of the data that are being analyzed.

Returning to the work of Fotheringham and Wong (1991) after assessing for the significance of the changes in parameter estimates, they investigated whether there was a link between these changes and spatial autocorrelation in the variables included in the analysis. Their conclusion was that there was little link between the severity of the MAUP and the degree of spatial autocorrelation in a (pair of) variable(s). They reinforced this conclusion by citing the examples of the percentage of black individuals and the percentage of home owners as displaying regression parameters that behaved very similarly under aggregation in terms of the significant change magnitude but that possessed very different spatial autocorrelation structures.

The work of Flowerdew and Green (1994) provides a way into understanding the properties of data with spatial autocorrelation. Using simulated data, they explore

the outcomes of multiple realizations of areal units at a given scale. The use of simulated data was important as it enabled them to analyze data with known spatial autocorrelation properties in comparison with real data where spatial autocorrelations are not known and may be impacted by other (unmeasured) biases as well. Green and Flowerdew aggregated their basic grid of raw simulated data into new areal units in three ways: (a) randomly; (b) systematically, based on the value of one of the simulated variables; and (c) spatially, by combining spatially contiguous blocks. The new zones that were constructed aspatially with random aggregation show no change in the subsequent correlation or regression outcomes (although the standard error is increased as a consequence of having fewer data points); the systematic aggregation increases the correlation coefficient but has no effect on the regression parameter, while spatial aggregation alters both coefficients. In conclusion, they argue that the effects of spatial autocorrelation may “result from contiguous processes affecting the distribution of one or more of the variables being analysed, or the spatial distribution of other variables which have effects on these.” This explicitly expresses the realization that the variables of areal units may display linked characteristics.

Developing their work on spatial autocorrelation further, Green and Flowerdew (1996) and Flowerdew and colleagues (2001) extend their analysis to consider the impact of spatial autocorrelation between variables as well as within variables, a phenomenon which they term “cross-correlation.” They define cross-correlation as the relationship not only between variable X and variable Y at a specific point in space but also being between X and Y at neighboring points in space. In Green and Flowerdew (1996), they continue using the simulated data but this time aggregated into spatially contiguous zones. They then model the relationship between the simulated X and Y firstly using a standard regression model and then using a model that incorporates the simulated cross-correlation between X and Y. Green and Flowerdew call the cross-correlation a regional effect, and they introduce a regional term into the regression model so that there is a regression coefficient for the local effect and a regression term for the regional effect. Having used simulated data for an initial exploration, attention is then turned to repeating the analysis with real data derived from the UK population census. Setting up an investigating into unemployment and ethnicity, Green and Flowerdew find evidence that confirms their cross-correlation hypothesis and demonstrates the usefulness of the local and regional regression approaches. In Flowerdew et al. (2001) they illustrate the same concept using the example from Fotheringham and Wong (1991, see above). They theorize that cross-correlation can occur because the relationship between the “attractiveness of housing (and hence its value and the likely income of the residents) may depend not just on race and class in the immediate vicinity but also on such characteristics in neighboring areas” (Flowerdew et al. 2001, p.91). Within this work is the useful conclusion that while the presence of spatial autocorrelation is important in determining the incidence of the scale effect in correlation coefficients, it does not impact on the regression coefficients. The regression coefficients are altered when cross-correlation is present between the X and Y variable.

Arbia (1989) introduced the term “systematic spatial variation” to create a formal framework to understand the relationship between the MAUP and spatial autocorrelation using Cliff and Ord’s work (1981) as a starting point. Using data relating to the residential location of population organized on a 32 by 32 lattice, Arbia simulated the MAUP by aggregating the grid into combinations of 16 by 16, 8 by 8, 4 by 4, and 2 by 2. The results of the investigation demonstrate that with aggregation there is an increase in the level of variance and that as the level of aggregation increases, the estimates of the variance of the data become more unreliable as the number of observations diminishes with fewer degrees of freedom. Arbia concluded the effects of the MAUP under aggregation were the result of the relationships between near objects. Building on this finding, Manley et al. (2006) demonstrate that spatial autocorrelation structures rarely match the boundaries of the zones that have been used to represent the data and that these differences between the spatial extent of the autocorrelation is, in part, one of the causes of the MAUP.

Over time, more complex models were applied to the MAUP. For instance, Amrhein and Flowerdew (1989) investigated the effects of MAUP in relation to Poisson regression. The results of their analysis demonstrated that within the Poisson model there is little zonation effect to be found. However, this is not a cause for celebration by the spatial analyst because a methodology to overcome the MAUP has been identified: the lack of effect is the consequence of the analytical technique, not because the results are free from the MAUP. The finding of Amrhein and Flowerdew is important because they add a new dimension to the MAUP discussion. They demonstrate that the choice of model for an analysis is just as critical as the zonation and scale choice itself. This conclusion does not, however, mean that the world of the analyst dealing with spatial data is bleak as might initially be presumed. Amrhein (1995) uses the finding above to develop six heuristics for analysts and suggest that certain statistics and results (for instance, the standard deviation of coefficients, or the Pearson correlation coefficient) exhibit greater changes due to MAUP (scale) than other statistical methods (for instance, mean or the variance).

The work investigating spatial autocorrelation, and the related cross-correlation, has demonstrated that the MAUP is likely to be caused by the interrelated nature of the spatial variables being represented in the areal units. Thus, when aggregation is undertaken and the spatial structure of the data has a direct influence on the resulting zonations the MAUP occurs. Manley et al. (2006) further demonstrated the complexity of this problem by analyzing British Census data and showing that spatial autocorrelation rarely coincides with the boundary lines of areal units and when aggregation is undertaken it frequently incorporates small zones with differing degrees of spatial autocorrelation.

59.3.3 Exploring the MAUP Through Zone Design

A cursory overview of the statistical investigations into the MAUP would suggest that the vast majority of effort into explaining the MAUP has been concerned with the scale effect. In fact, the zonation issue has also been tackled extensively, and in

some regards, with more success than the scale issue. The zonation issue research has largely focused on two aspects: how can zonations be created that are appropriate to the analytical task and what are the properties of zonation that lead to the MAUP occurring. The ability to provide multiple realizations of zonal systems within one analysis space enables the scale effect to be investigated further, as many different zonations can be derived as scale changes.

If zoning systems are problematic, then it is useful to consider why and how zoning systems may be (re)designed. The rationale behind is summed up by Openshaw and Rao (1995): “[t]he new opportunity provided by [the increasing availability of digital] boundaries is not to demonstrate the universality of MAUP effects, or to manipulate results by gerrymandering the spatial aggregation used, but it is to design new zoning systems that may help users recover from MAUP.” Openshaw (1978) presented two extremes of zone design approaches to illustrate the problem. A conventional statistical approach within which spatially aggregated data can be viewed as fixed, or a model that assumes that the “undefined parameters [are] fixed, and the identification of an appropriate zoning system has to be made in some optimal manner.” The first view is unacceptable due to the interdependence between the choice of zone and results achieved. From a statistical standpoint, the second solution is as poor as the first one was from a geographic perspective, as it could serve to remove the comparability between studies.

The process of zone design presents a compromise through the creation of the system that satisfies (or at least suffices) a set of criteria. One ideal outcome for a good zonal system would be a set of zones that was as simple as possible, homogenous (against a single or set of variables defined by the user), and compact. In contrast, Openshaw (1978) increased the complexity of the problem and suggested that shape (as distinct to compactness) and population size are also important elements to include. Depending on the task for which the zones are required, each of these criteria may be made more or less important. One of the first attempts at automated zone design was undertaken by Stan Openshaw (1978) with the Automatic Zoning Procedure, implemented in the Automatic Zoning Program (AZP). In more recent research, the process of zone design has become integrated with the mainstream literature around Geographical Information Systems (GIS) and enabled users to define their own zonal units. AZP was extended and became the Zone Design System (ZDES) and has been employed in a wide range of zonal scenarios. One prime example is explored in Openshaw and colleagues (1998) which commented on the first fully automated basic spatial unit (bsu) design process undertaken for the publication of the 2001 UK Census data. As Openshaw and colleagues point out, one of the major barriers to successful zone design is the realization that the problem is not one that can be tackled in the traditional software programming sense, where a global optimal solution is identifiable – if there was a global optimal solution, it is not clear how it would be identified, and in many cases there is no optimal solution. Rather, there is a range of suitable solutions which present sufficient solutions given the criteria that have been inputted.

Other systems have been developed specifically for zonal data analysis and redesign. An alternative to ZDES, is AZM (Automated Zone Matching). AZM “[i]mplements zone design on a set of zones described by polygon and arc attribute tables exported from Arc/Info or generated by users’ own programs. [The program is designed to optimize] the match between two zonal systems, or the aggregation of a set of building block zones into output areas with a range of user-controlled design parameters” (Martin 2003). AZM uses the AZP procedure outlined by Openshaw (1978) and is conceptually similar. However, unlike ZDES, the AZM program was not designed specifically for the purpose of zone design. The primary function of the program is to provide a means to enable two incompatible zone coverages to be aggregated into a higher-level zone system that enables comparison (Martin 2003). However, through the input of two identical coverages, it can be used to perform an aggregation function (Martin (2003)). Nevertheless, the advantages of being able to control the aggregation process with regard to shape, key variable homogeneity, and population size mean that it is suited to the design of analytically appropriate zonal systems. In other words, zonal systems that better reflect the required uses of data, as opposed to purely “random” aggregations where there is little or no control over one or all of these factors, are not relevant in the context of research where desired scales of aggregation are required.

Finally, evidence of the potency of understanding zone design and exploiting it was presented by Boyle and Alvanides (2004). Using a case study involving the City of Leeds, and measures of deprivation, they demonstrate that it is possible to change the ranking of Leeds relative to other cities across the UK by using different boundary systems. This is of particular importance, as the European Commission was offering what are termed structural funds to aid the reduction of inequalities at a local level within member countries. Using the 1998 Index of Local Deprivation (ILD) based on the 1991 Census, as published, Leeds appeared 56th out of 57 cities. However, simply by redrawing the boundaries using alternative population thresholds to define the city area, the ranking could be changed to 11th. Applying another different criteria for the aggregation, whereby the scores were taken for wards, not local authority districts, enabled a further change in the ranking, making Leeds the 3rd most deprived city in England. The initial ranking of 56th would not have secured funding while the final ranking of 3rd would ensure a large flow of money into the city. Both of these examples highlight the potential difficulties, opportunities, and concerns that research using aggregated data should address.

59.4 Conclusion

This chapter has provided an overview of the modifiable areal unit problem (MAUP). With the growth of spatially coded data available, the potential for analysts to be confronted with areal units in analysis is increasing dramatically. Knowledge of the potential pitfalls of conducting analysis containing areal unit data is vital when dealing with areal unit data in analysis. This is true both when the areal units are the objects of the analysis as it is when the areal unit data are included to

provide context to other sorts of information. In many cases, it is important to acknowledge the presence of the MAUP in analysis while accepting that the results may be conditional on the scale and zonation scheme employed.

Previous research has demonstrated that it is unlikely that a global solution to the MAUP will ever be found: indeed, to do so is to deny the inherent spatiality of the data that is under investigation, and the removal of the MAUP would be to remove the very object of interest! Previous research has also demonstrated that spatial autocorrelation and cross-correlation are likely to be very important in understanding the degree and severity of the MAUP. As such, these are key topics that the (spatial) analyst using aggregate data should be aware of and acknowledge in their analysis. Therefore, when dealing with spatially organized data, the analyst must adopt a geographically informed process of hypothesis formation. Analytical scale should become a primary factor that is explicitly considered rather than an issue that is implicitly dealt with and all too frequently assumed away in the name of pragmatism. In many cases, this will require the analyst to adopt an approach whereby multiple scales of measurement and analysis should be considered, or a highly rigorous spatial framework for an analysis constructed. This chapter is all too brief to provide a comprehensive view of all the work that has been conducted into the MAUP. Nevertheless, it hopefully sheds sufficient light on the subject and processes to provide the reader with the means to adopt a more critical and nuanced approach to their analysis.

References

- Amrhein C (1995) Searching for the elusive aggregation effect: evidence from statistical simulations. *Environ Plann A* 27(1):105–119
- Amrhein C, Flowerdew R (1989) The effect of data aggregation on a Poisson regression model of Canadian migration. In: Goodchild MF, Gopal S (eds) Accuracy of spatial databases. Taylor and Francis, London, pp 229–238
- Arbia G (1989) Spatial data configuration in statistical analysis of regional economic and related problems. Kluwer, Dordrecht
- Baumann J, Fischer MM, Schubert U (1983) A multiregional labour supply model for Austria: the effects of different regionalisations in labour market modelling. *Pap Reg Sci Assoc* 52(1):214–218
- Boyle P, Alvanides S (2004) Assessing deprivation in English inner city areas: making the case for EC funding for Leeds city. In: Clarke G, Stillwell J (eds) Applied GIS and spatial analysis. Wiley, Chichester, pp 111–136
- Cliff AD, Ord JK (1981) Spatial processes: models & applications. Pion, London
- Duncan OD, Duncan B (1955) A methodological analysis of segregation indices. *Am Sociol Rev* 20:210–217
- Flowerdew R (2011) How serious is the modifiable areal unit problem for analysis of English census data? *Population Trends* 145 (Autumn). Office for National Statistics, pp 1–13
- Flowerdew R, Geddes A, Green M (2001) Behaviour of regression models under random aggregation. In: Tate NJ, Atkinson PM (eds) Modelling scale in geographical information science. Wiley, Chichester, pp 89–104
- Flowerdew R, Green M (1994) Areal interpolation and types of data. In: Fotheringham S, Rogerson P (eds) Spatial analysis and GIS. Taylor and Francis, London, pp 121–145

- Fotheringham AS, Wong DWS (1991) The modifiable areal unit problem in multivariate statistical analysis. *Environ Plann A* 23(7):1025–1044
- Gehlke CE, Biehl K (1934) Certain effects of grouping upon the size of the correlation in census tract material. *J Am Stat Assoc* 29(Special Suppl):169–170
- Getis A (2010) Spatial autocorrelation. In: Fischer MM, Getis A (eds) *Handbook of applied spatial analysis. Software, methods and applications*. Springer, Berlin/Heidelberg/New York, pp 255–278
- Green M, Flowerdew R (1996) New evidence on the modifiable areal unit problem. In: Longley P, Batty M (eds) *Spatial analysis: modelling in a GIS environment*. GeoInformation International, Cambridge, pp 41–54
- Heasman MA, Kemp W, McLaren AM, Trotter P, Gillis CR, Hole DJ (1984) Incidence of leukaemia in young persons in west of Scotland. *Lancet* 323(8387):1188–1189
- Kirby AM, Taylor PJ (1976) A geographical analysis of voting patterns in the EEC referendum. *Regional Stud* 10(2):183–191
- Manley D, Flowerdew R, Steel D (2006) Scales, levels and processes: studying spatial patterns of British census variables computers. *Environ Urban Syst* 30(1):143–160
- Martin D (2003) Developing the automated zoning procedure to reconcile incompatible zoning systems. *Int J Geogr Inform Sci* 17(1):181–196
- Odoi A, Martin SW, Michel P, Holt J, Middleton D, Wilson J (2003) Geographical and temporal distribution of human giardiasis in Ontario, Canada. *Int J Health Geogr* 2(1):5
- Openshaw S (1977) A geographical solution to the scale and aggregation problem in region-building, partitioning and spatial modeling. *Trans Inst Br Geographr, New Ser* 2(4):459–472
- Openshaw S (1978) An empirical study of some zone-design criteria. *Environ Plann A* 10(7):781–794
- Openshaw S (1984) The modifiable areal unit problem. CATMOG 38. GeoBooks, Norwich
- Openshaw S, Rao L (1995) Algorithms for reengineering 1991 census geography. *Environ Plann A* 27(3):425–446
- Openshaw S, Taylor PJ (1979) A million or so correlation coefficients, three experiments on the modifiable areal unit problem. In: Wrigley N (ed) *Statistical applications in the spatial sciences*. Pion, London, pp 127–144
- Openshaw S, Taylor PJ (1981) The modifiable areal unit problem. In: Bennet RJ, Wrigley N (eds) *Quantitative geography*. Routledge Kegan Paul, Henley-on-Thames, pp 60–69
- Openshaw S, Alvanides S, Whalley S (1998) Some further experiments with designing output areas for the 2001 UK census. In: The paper presented at the 4th of the ESRC/JISC supported workshops Planning for the 2001 Census
- Poulsen M, Johnston R, Forrest J (2011) Using local statistics and neighbourhood classifications to portray ethnic residential segregation: a London example. *Environ Plann B* 38(4):636–658
- Tobler W (1970) A computer movie simulating urban growth in the Detroit region. *Econ Geogr* 46(2):234–240
- Wilkie D (1986) Precognition on: review of the Scottish Health Service ISD report on Geographical distribution of leukaemia in young persons in Scotland 1968–1983 (document D/P/20) (Presented at EDRP public local inquiry, Thurso, September, 1986)
- Wong D (2003) Spatial decomposition of segregation indices: a framework toward measuring segregation at multiple levels. *Geograph Anal* 35(3):179–194
- Wong D (2009) Chapter 7. The modifiable areal unit problem (MAUP). In: Fotheringham AS (ed) *The SAGE handbook of spatial analysis*. Springer, Dordrecht, pp 95–112
- Yule GU, Kendal MG (1950) An introduction to the theory of statistics. Charles Griffin and Company Limited, London

Tao Cheng, James Haworth, Berk Anbaroglu,
Garavig Tanaksaranond, and Jiaqiu Wang

Contents

60.1	Introduction	1173
60.2	Spatio-Temporal Autocorrelation	1176
60.2.1	The Global Measure	1177
60.2.2	The Local Measure	1177
60.3	Space-Time Forecasting and Prediction	1178
60.3.1	Statistical (Parametric) Models	1178
60.3.2	Machine Learning (Non-parametric) Approaches	1181
60.3.3	Summary	1183
60.4	Space-Time Clustering	1183
60.4.1	Introduction	1183
60.4.2	Spatio-Temporal Scan Statistics	1184
60.5	Space-Time Visualization	1185
60.5.1	2D maps	1185
60.5.2	3D Visualization	1188
60.5.3	Animated Maps	1188
60.5.4	Visual Analytics: The Current Visualization Trend	1189
60.6	Conclusions	1190
	References	1192

Abstract

As the number, volume and resolution of spatio-temporal datasets increases, traditional statistical methods for dealing with such data are becoming overwhelmed. Nevertheless, the spatio-temporal data are rich sources of

T. Cheng (✉) • J. Haworth • B. Anbaroglu • G. Tanaksaranond • J. Wang
SpaceTimeLab, Department of Civil, Environmental and Geomatic Engineering, University College London, London, UK
e-mail: tao.cheng@ucl.ac.uk; j.haworth@ucl.ac.uk; b.anbaroglu@ucl.ac.uk;
g.tanaksaranond@ucl.ac.uk; w.jiaqiu@ucl.ac.uk

information and knowledge, waiting to be discovered. The field of spatio-temporal data mining (STDM) emerged out of a need to create effective and efficient techniques in order to turn the massive data into meaningful information and knowledge. This chapter reviews the state of the art in STDM research and applications, with emphasis placed on three key areas, including spatio-temporal prediction and forecasting, spatio-temporal clustering and spatio-temporal visualization. The future direction and research challenges of STDM are discussed at the end of this chapter.

60.1 Introduction

With automatic sensor networks and crowd sourcing now being used extensively to monitor a diverse range of phenomena, the amount of data being collected with both spatial and temporal dimensions has increased dramatically. Data collected at two or more locations and times make up space-time series, examples of which include daily temperature series at meteorological stations, monthly crime rates of world capital cities and daily traffic flow on urban roads. These space-time series are massive and continually growing. Spatio-temporal data mining (STDM) is the extraction of unknown and implicit knowledge, structures, relationships, or patterns from these massive datasets. STDM techniques and tasks include spatio-temporal forecasting, spatio-temporal association rule mining, spatio-temporal sequential pattern mining and spatio-temporal clustering and classification, amongst others (Miller and Han 2009). More recently, spatio-temporal visualization has become another hot topic for STDM as we begin to explore new ways of representing spatio-temporal data that go beyond the static map.

Early research efforts on spatio-temporal forecasting focused on adapting existing statistical regression models from the fields of time series analysis, spatial analysis and econometrics to deal with spatio-temporal data. Such models are typically geared towards teasing scarce information from homogenous datasets and have been overwhelmed by the increasing volume and diversity of spatio-temporal data that is now being collected. Increasingly, researchers and practitioners are turning towards less conventional techniques, often with their roots in the machine learning and data mining communities, that are better equipped to deal with the heterogeneous, nonlinear and multi-scale properties of large scale spatio-temporal datasets. For instance, methods such as artificial neural networks (ANNs) and support vector machines (SVMs) are now being successfully applied to spatio-temporal forecasting problems.

The association (or co-location) rule mining is to infer the presence of spatial features in the neighbourhood of other spatial features (Shekhar et al. 2011). They are spatial extensions of association rules, which were developed by the retail industry to examine the behavior of consumers. A spatio-temporal co-location rule implies a strong association between locations A and B that if the attributes of A take some specific value at a point in time, then with a certain probability, at the same point in time, the attributes of B will take some specific value. A related

STDM task is mixed drove co-occurrence pattern (MDCOP) mining. MDCOPs are subsets of two or more different object types whose instances are often located close to one another in space and time (Shekhar et al. 2011). The drawback of these methods is that only contemporaneous associations are considered so they do not account for the evolution of a spatial process over time.

A logical extension to association mining is to analyze spatio-temporal sequential patterns. This involves finding sequences of events (an ordered list of item sets) that occur frequently in spatio-temporal datasets. Sequential pattern mining algorithms were also first introduced to extract patterns from customer transaction databases. A spatio-temporal sequential pattern means that if at some point in time and space, the attributes in A take some specific value, then with a certain probability at some later point in time, attributes at B will take some specific value. Sequential pattern mining implicitly incorporates the notion of spatio-temporal dependence; that the events at one location at one time can have some causal influence on the events at another location at a subsequent time. A similar concept to sequential patterns are cascading spatio-temporal patterns, which are ordered subsets of events that are located close together and occur in a cascading sequence (Shekhar et al. 2011).

Clustering involves grouping unlabeled objects that share similar characteristics. The goal is to maximize the intraclass similarity and minimize the interclass similarity. Clustering can be used for classification, segmentation and outlier detection, and here clustering is a general term for all these tasks. Widely used spatial clustering techniques e.g., K-means and K-medoids, have been extended to spatio-temporal clustering problems. Designing an effective spatio-temporal clustering algorithm is a difficult task because it must account for the dynamics of a phenomenon in space and time. For instance, when clustering moving objects, a cluster may change its spatial location from one time step to the next but still be the same spatio-temporal cluster. Rules for capturing this type of behavior are difficult to encode in algorithms.

Mining interesting patterns, rules and structures from spatio-temporal data is only part of the task of STDM. The results are not useful if they are not easily understood. For instance, finding a spatio-temporal cluster in a patient register dataset is not useful in itself. On the other hand, confirming this spatio-temporal cluster as a disease outbreak and visualizing it using a platform that epidemiologists and medical professionals can understand is very useful indeed. As a result, space-time visualization has emerged as another important facet of STDM. It explores the patterns hidden in the large data sets by using advanced visualization and animation techniques. This includes conventional 2D maps as well as newly developed 3D space-time cube methods, which can show hotspots and isosurfaces of spatio-temporal phenomena. Integration of data exploration, analysis and visualization in a single platform takes this one step further. The STARS platform (space-time analysis of regional systems, Rey and Janikas 2010) is an excellent example of this that allows exploratory and explanatory analysis and visualization of regional data with spatio-temporal extent. However, despite significant progress, how to visualize large volumes of data in real time and to best make use of the third dimension are problems that are yet to be adequately solved.

This chapter is organized around three main tasks of STDM; space-time modeling and prediction, space-time clustering and space-time visualization. In the following section, we review spatio-temporal autocorrelation and its implications for space-time modeling. [Section 60.3](#) is devoted to space-time modeling and prediction, by either statistical (parametric) approaches or machine learning (non-parametric) approaches. [Section 60.4](#) gives a brief review of space-time clustering and outlier detection, and is followed by an introduction to space-time visualization in [Sect. 60.5](#). The final section summarizes the directions of future research in STDM.

60.2 Spatio-Temporal Autocorrelation

An observation from nature is that near things tend to be more similar than distant things both in space and in time. For instance, the weather tomorrow is more likely to be similar to today's weather than the weather a week ago, or a month ago and so on. Similarly the weather 1 mile away is likely to be more similar than the weather 10 miles away or 100 miles away. These phenomena are referred to respectively as temporal and spatial dependence. The presence of dependence in spatial and temporal data violates the stationarity assumption of classic statistical models such as ordinary least squares (OLS) and necessitates the use of specialized modeling and forecasting techniques. Testing for dependence is typically accomplished using an autocorrelation analysis. Autocorrelation is the cross-correlation of a signal with itself and can be measured in temporal data using the temporal autocorrelation function (ACF, [Box and Jenkins 1970](#)) or in spatial data using an index such as the familiar Moran coefficient.

These measures are global, implying a degree of fixity in the level of autocorrelation across the space/time such that it can be described by a single parameter. However, this is often unrealistic. Many time series exhibit nonlinear characteristics that make stationarization difficult. Similarly, spatial data often exhibit structural instability over space, which is referred to as heterogeneity. Heterogeneity has two distinct aspects; structural instability as expressed by changing functional forms or varying parameters, and heteroskedasticity that leads to error terms with non-constant variance ([Anselin 1988](#)). Ignoring it can have serious consequences including biased parameter estimates, misleading significance levels and poor predictive power. [Anselin \(1988\)](#) provides some methods for testing for heterogeneity. Additionally, a number of local indicators of spatial association (LISA) have been devised. These include a local variant of Moran's I and Getis and Ord's G_i and $G_{\bar{i}}$ statistics, which measure the extent to which high and low values are clustered together.

Although sharing many commonalities in techniques and concepts, the fields of time series analysis and spatial analysis have largely developed separately from one another. The behavior of a variable over space differs from its behavior in time. Time has a clear ordering of past, present and future while space does not and because of this ordering isotropy has no meaning in the space-time context. In time, measurements can only be taken on one side of the axis; hence estimation involves

extrapolation rather than interpolation. Temporal data also has other characteristics, such as periodicity, that are not common in spatial data and scales of measurement also differ between space and time and are not directly comparable.

When a variable Z is observed over time at two or more locations, it is both a spatial series and a time series and can be referred to as a space-time series $z = \{z(s, t) | s \in S, t \in T\}$ in spatial domain S and temporal interval T . A space-time series may exhibit spatio-temporal dependence which describes its evolution over space and time. If the spatio-temporal dependence in a dataset can be modeled then one essentially has predictive information. A number of indices have been devised to this end including space-time (semi) variograms (Heuvelink and Griffith 2010) as well as space-time eigenvector filtering (Griffith 2010). Two indices are described here, the space-time autocorrelation function (ST-ACF), that measures global space-time autocorrelation, and the cross-correlation function (CCF), that measures local space-time autocorrelation between two locations. These indices are extensions of the temporal autocorrelation function and are selected as they are easily interpretable and have a practical application in established space-time modeling frameworks.

60.2.1 The Global Measure

The ST-ACF measures the N^2 cross-covariances between all possible pairs of locations lagged in both time and space (Pfeifer and Deutsch 1980). Given the weighted l^{th} order spatial neighbours of any spatial location at time t and the weighted K^{th} order spatial neighbors of the same spatial location s time lags in the future, the space-time cross-covariance can be given as:

$$\gamma_{lk}(s) = E \left\{ \frac{[W^{(l)}z(t)]' [W^{(k)}z(t+s)]}{N} \right\} \quad (60.1)$$

Where N is the number of spatial locations, $W^{(l)}$ and $W^{(k)}$ are the $N * N$ spatial weight matrices at spatial orders l and k , $Z(t)$ is the $N * 1$ vector of observations z at time t , $z(t+s)$ is the $N * 1$ vector of observations z at time $(t+s)$ and the symbol $'$ denotes matrix transposition. Based on Eq. (60.1), the ST-ACF can be defined as:

$$\rho_{lk}(s) = \frac{\gamma_{lk}(s)}{[\gamma_{ll}(0)\gamma_{kk}(0)]^{\frac{1}{2}}} \quad (60.2)$$

ST-ACF has been used in STARIMA to calibrate the order of moving average (MA), which define the range of spatial neighbourhoods which contribute to the current location at a specific time lag (Pfeifer and Deutsch 1980). The MA orders are fixed globally both spatially and temporally and a single parameter is estimated for it in practical application such as in Kamarianakis and Prastacos (2005), and Cheng et al. (2011b).

60.2.2 The Local Measure

The cross correlation function (CCF) (see, for example, Box and Jenkins 1970) treats two time series as a bivariate stochastic process and measures the cross covariance coefficients between each series at specified lags. It provides a measure of the similarity between two time series. The CCF is useful if one has reason to believe that the level of autocorrelation in a spatio-temporal dataset is not fixed in time and space. Given two time series X and Y, the CCF at lag k is given as:

$$\rho_{xy}(k) = \frac{E[(x_t - \mu_x)(y_{t+k} - \mu_y)]}{\sigma_x \sigma_y} \quad k = 0, \pm 1, \pm 2, \pm \dots \quad (60.3)$$

The CCF measures cross-correlations in both directions, as denoted by subscript k, therefore the temporal lag at which the CCF peaks can be used to determine a transfer function between two series. This is, however, dependent on sufficient spatial and temporal resolution in the data. A peak at lag zero indicates that the current resolution does not capture the direction of influence of one location on another, but the series behave very similarly at the same time (Cheng et al. 2011a). As examples, the global and local measures of road network in central London are shown in Figs. 60.1 and 60.2.

60.3 Space-Time Forecasting and Prediction

Space-time models must account for the combined problems of spatial and temporal data mentioned in the preceding sections. Uptake of space-time models has traditionally been limited by the scarcity of large scale spatio-temporal datasets (Griffith 2010). This is a situation that has been reversed over recent decades and we are now inundated with data and require methods to deal with them quickly and effectively. The models that are currently applied to space-time data can be broadly divided into two categories; statistical (parametric) methods and machine learning (non-parametric) methods. These are described in turn in the following subsections.

60.3.1 Statistical (Parametric) Models

The state of the art in statistical modeling of spatio-temporal processes represents the outcome of several decades of cross-pollination of research between the fields of time series analysis, spatial statistics and econometrics. Some of the methods commonly used in the literature include space-time autoregressive integrated moving average (STARIMA) models (Pfeifer and Deutsch 1980) and variants, multiple ARIMA models, space-time geostatistical models (Heuvelink and Griffith 2010), spatial panel data models (Elhorst 2003), geographically and temporally weighted regression (Huang et al. 2010) and eigenvector spatial filtering (Griffith 2010).

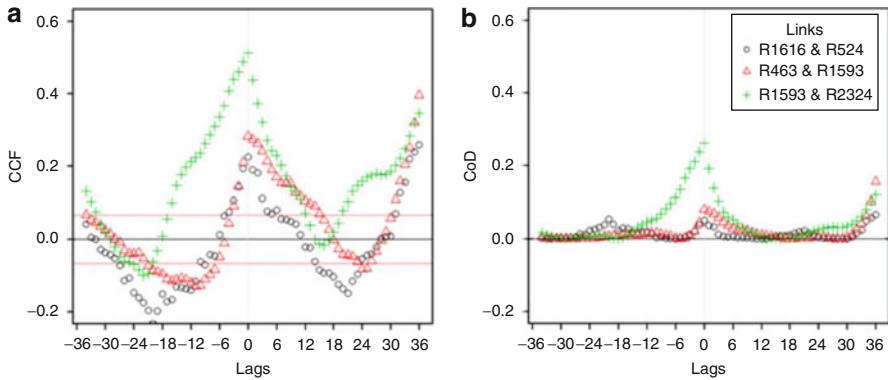


Fig. 60.1 (a) CCF and (b) coefficient of determination (CCF^2) between unit journey times of three pairs of road links in central London in the AM peak period (7–10am) (Cheng et al. 2011a)

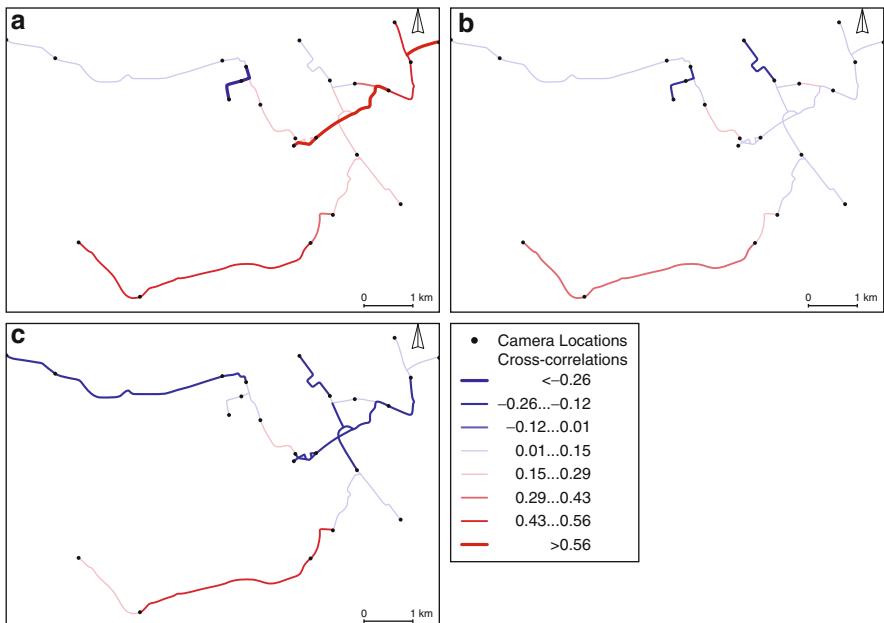


Fig. 60.2 Average CCF between links and their first-order neighbours at temporal lag zero in (a) the AM peak; (b) interpeak; and (c) PM peak (Cheng et al. 2011a)

60.3.1.1 Space-Time Autoregressive Integrated Moving Average

Space-time autoregressive integrated moving average (STARIMA) is a family of models that extend the ARIMA time series model to space-time data (Pfeifer and Deutsch 1980). STARIMA explicitly takes into account the spatial structure in the data through the use of a spatial weight matrix. The general STARIMA model

expresses an observation of a spatial process as a weighted linear combination of past observations and errors lagged in both space and time. A fitted STARIMA model is usually described as a STARIMA (p,d,q) model, where p indicates the autoregressive order, d is the order of differencing and q is the moving average order. The application of STARIMA models has been fairly limited in the literature, with examples existing in traffic prediction (Kamarianakis and Prastacos 2005) and temperature forecasting (Cheng et al. 2011b).

Some important special cases of the STARIMA model should be noted; when $d = 0$ the model reduces to a STARMA model, furthermore, a STARMA model with $q = 0$ is a STAR model and with $p = 0$ is a STMA model. Although the STARIMA model family accounts for spatio-temporal autocorrelation, it has not yet been adequately adapted to deal with spatial heterogeneity and parameter estimates are global. The implication of this is that the space-time process must be stationary (or made stationary through differencing/transformation) for STARIMA modeling to be effective.

60.3.1.2 Spatial Panel Data Models

Panel data is a term used in the econometrics literature for multi-dimensional data. A panel contains observations on multiple phenomena (cross-sections) over multiple time periods. When panel data include a spatial component they are referred to as spatial panel data. Although the term describes the data itself, there are a range of models that have been developed to work with spatial panel data that originate specifically from spatial econometrics that are referred to as spatial panel data models. Methodologically, they are often very similar to those encountered in the spatial statistics literature.

Aspatial panel data models are modified to account for spatial dependence in one of two ways; either with a spatial autoregressive process in the error term; a spatial error model (equivalent to a spatial moving average), or with a spatially autoregressive dependent variable; a spatial lag model (Elhorst 2003). In their standard form, spatial panel data models are global models and do not account for spatial heterogeneity and, as in the spatial statistics literature, this has become a focus of research in recent years. Elhorst (2003) defined a set of spatial panel data models that account for heterogeneity in different ways. The uptake of spatial panel data models has been much more widespread than those mentioned in Sect. 60.3.1 and there have been applications in liquor demand prediction and US state tax competition, amongst many others.

60.3.1.3 Space-Time GWR

Recently, there has been a great deal of interest in extending geographically weighted regression (GWR) to the temporal dimension. In their geographically and temporally weighted regression (GTWR) model, Huang et al. (2010) incorporate both the spatial and temporal dimensions into the weight matrix to account for spatial and temporal nonstationarity. The technique was applied to a case study of residential housing sales in the city of Calgary from 2002 to 2004 and found to outperform GWR and temporally weighted regression (TWR) as well as OLS.

60.3.1.4 Space-Time Geostatistics

Space-time geostatistics is concerned with deriving space-time covariance structures and semivariograms for the purpose of space-time interpolation and forecasting. The aim is to build a process that mimics some patterns of the observed spatiotemporal variability, without necessarily following the underlying governing equations (Kyriakidis and Journel 1999). The first step usually involves separating the deterministic component $m(u, t)$ of space time coordinates u and t . Following this, a covariance structure is fitted to the residuals. The simplest approach is to separate space and time and consider the space-time covariance to be either a sum (zonal anisotropy model) or product (separable model) of separate spatial and temporal covariance functions. Although simple to implement, these models have the disadvantage that they do not consider space-time interaction. They assume a fixed temporal pattern across locations and a fixed spatial pattern across time. Additionally, it is not straightforward to separate the component structures from the experimental covariances. For example, an experimental spatial covariance will be influenced by temporal variability resulting from the time instant at which the data was measured.

The second approach is to model a joint space-time covariance structure. This approach is generally accepted to be more appropriate. Combinations of the two approaches have also been described in the literature (Heuvelink and Griffith 2010). Once an appropriate space-time covariance structure has been defined, one can use standard Kriging techniques for interpolation and prediction; Space-time geostatistical techniques are best applied to stationary space-time processes. Highly nonstationary spatio-temporal relationships require a very complicated space-time covariance structure to be modelled for accurate prediction to be possible. Despite being spatio-temporal in nature, the main function of space-time geostatistical models is space-time interpolation and they encounter problems in forecasting scenarios where extrapolation is required (Heuvelink and Griffith 2010).

60.3.2 Machine Learning (Non-parametric) Approaches

In parallel to the development of statistical space-time models, there was a multidisciplinary explosion of interest in non-parametric machine learning methods, and many of these have been successfully adapted to work with spatio-temporal data due to their innate ability to model complex nonlinear relationships. There is a wide range of machine learning algorithms available, in this section we focus on two of the most popular; the artificial neural network and the support vector machine.

60.3.2.1 Artificial Neural Networks

Artificial Neural Networks (ANNs) are a family of non-parametric methods for function approximation that have been shown to be very powerful tools in many application domains (see Fischer 2006 for example), often dealing with complex real world sensor data. They were initially inspired by the observation that biological learning is governed by a complex set of interconnected neurons. The key

concept is that, although individual neurons may be simple in structure, their interconnections allow them to perform complex tasks such as pattern recognition and classification.

Since its inception, the term ANN has become an umbrella term for a broad class of flexible non-linear models for regression and classification with a range of different architectures. ANNs have been widely applied in spatial and temporal analysis. Kanevski et al. (2009) have applied various types of ANN to spatial and environmental modeling problems including radial basis function neural networks (RBFNN), general regression neural networks (GRNN), probabilistic neural networks (PNN) and neural network residual Kriging (NNRK) models and have gained excellent results. The authors note that the strength of ANNs is that they learn from empirical data and can be used in cases where the modeled phenomena are hidden, non-evident or not very well described. This makes them particularly useful in modeling the complex dependency structures present in space-time data that cannot be described theoretically. Hsieh (2009) also provides a good review of ANN methods applied to spatial problems.

60.3.2.2 Support Vector Machines

Another widely used machine learning technique is the support vector machine (SVM, SVR in the regression case). SVMs are a set of supervised learning methods originally devised for classification tasks that are based on the principles of statistical learning theory (Vapnik 1999). SVMs make use of a hypothesis space of linear functions in a high dimensional feature space, trained with a learning algorithm from optimization theory. The key to their strong performance is that the learning task is formulated as a convex optimization problem meaning that, for a given set of parameters, the solution is globally optimal provided one can be found. Therefore, SVMs avoid the problem of getting stuck in local minima which are traditionally associated with ANNs. This has led to SVMs outperforming most other systems in a wide variety of applications within a few years of their introduction.

SVMs have been successfully used to model time series in a number of application areas including financial time series and traffic flow prediction. Compared to time series analysis, the uptake of SVM in the spatial sciences was initially slow but has seen a rapid increase in popularity in the past 5 years or so. The book “Machine Learning for Spatial Environmental Data” (Kanevski et al. 2009) provides a good introduction to some of the machine learning methods currently being used to model spatial data. Recently, SVMs have been applied to spatio-temporal avalanche forecasting (Pozdnoukhov et al. 2011). The approach involves incorporating the outputs of simple physics based and statistical approaches to interpolate meteorological and snowpack related data over a digital elevation model of the region. The decision boundary is used to discriminate between safe and dangerous conditions.

60.3.2.3 Other Methods

ANNs and SVMs are two methods that are widespread in temporal and spatial analysis, however, the field of machine learning is huge and a comprehensive

review is beyond the scope of this chapter. Various other methods have been used including nearest neighbour regression, kernel (ridge) regression, Gaussian processes, self-organizing maps (SOM), principal components analysis (PCA) and regression trees, which are introduced in Hsieh (2009). This list is non-exhaustive and readers are also referred to the text of Kanevski et al. (2009) for detailed introduction in the context of spatial data.

60.3.3 Summary

In this section, the complex, nonlinear, non-stationary properties of spatio-temporal data and their implications for space-time models were outlined. The question is which model should one choose for a given spatio-temporal dataset? The answer to this depends on the data. In the literature, space-time analysis is typically applied to data with low spatial and/or temporal resolution which is acquired after the event. In the tradition of spatial analysis, the practical use of such data is to elicit causal relationships between variables that can give some valuable insights into the underlying processes. In this case, the use of parametric statistical models may be preferable because of their explanatory power and interpretability.

However, these days, more and more data sources are becoming available in (near) real time at high spatial and temporal resolutions. Extracting meaningful relationships from such data is a task that is secondary to forecasting and it is likely that machine learning approaches, with their greater flexibility, will play an ever increasing role. Generally, machine learning methods have a wider field of application than traditional geostatistics due to their ability to deal with multi-dimensional nonlinear data. They are also well suited to dealing with large databases and long periods of observation. In particular, the SVM approach is favorable because it avoids the curse of dimensionality faced by other methods. One of the future research directions in this area lies in improving the interpretability of the structure and output of machine learning algorithms. Another way is to use a hybrid framework with both statistical and machine learning approaches (Cheng et al. 2011b).

60.4 Space-Time Clustering

60.4.1 Introduction

Another very important task of STDM is to extract meaningful patterns and relationships from massive spatio-temporal data that are not necessarily explicit. In this situation, we may wish to search for structure in the dataset without an *a priori* hypothesis. Hypotheses can be then be formed and refined *aposteriori* from the results. This is known as unsupervised learning. One of the most important unsupervised learning tasks in STDM is clustering. This involves grouping space-time series into clusters, where the similarity of data within a cluster and the dissimilarity between the clusters are high. Clustering can also be used to detect outliers.

A spatial outlier is a spatially referenced object whose thematic attribute values are significantly different from those of other spatially referenced objects in its spatial neighborhood. It represents an object that is significantly different from its neighborhoods even though it may not be significantly different from the entire population. A spatial-temporal outlier is a spatial-temporal object whose thematic attribute values are significantly different from those of other spatially and temporally referenced objects in its spatial or/and temporal neighborhoods. Identification of ST-outliers can lead to the discovery of unexpected, interesting, and implicit knowledge, such as local instability or deformation (Cheng and Li 2006). Nowadays spatial and spatio-temporal clustering has been widely used to understand the spatial patterns hidden in spatial databases with applications in epidemic studies, crime hotspot analysis and social networks.

The three domains of space-time series data can be used to define the similarity between observations for clustering. The first is the thematic domain, where the attributes define the characteristics of the object. The second is the spatial domain, which is used to describe the location of the object. Finally, the temporal domain is used to store the timing of the object. These domains are used to answer the questions *what*, *where* and *when* respectively. Initial research on clustering focused on the thematic domain, with methods such as k-means, k-medoids and their variants being popular. Research into clustering using the spatial domain gained popularity in the early twenty-first century. Initial research on spatial clustering has focused on point data. Popular algorithms such as DBSCAN and BIRCH are the outputs of this research area. The spatial distance or the spatial density derived from the spatial locations of the points is considered for clustering. Clustering has also been conducted by combining spatial adjacency with thematic domains or by combining the spatial distance with the thematic distance. Temporal-thematic clustering is mainly applied to group time series data, in order to know whether customers are changing over time, or to determine if credit card fraud transactions change over time.

Very few algorithms consider the spatial, temporal and thematic attributes seamlessly and simultaneously in the clustering. Capturing the dynamicity in the data is the most difficult challenge in spatio-temporal clustering, which is the reason that traditional clustering algorithms, in which the clustering is carried out on a cross-section of the phenomenon, cannot be directly applied to spatio-temporal phenomena. The arbitrarily chosen temporal intervals may not capture the real dynamics of the phenomena since they only consider the thematic values at the same time, which cannot capture the influence of flow (i.e., time-lag phenomena). It is only recently that this has been attempted. We pay particular attention to spatio-temporal scan statistics, a method that has shown promising performance in a range of STDM tasks such as health, crime and transport studies.

60.4.2 Spatio-Temporal Scan Statistics

Spatio-temporal scan statistics (STSS) is a clustering technique that was originally devised to detect disease outbreaks (Neill 2008). The goal is to automatically detect

regions of space that are “anomalous,” “unexpected,” or otherwise “interesting.” Spatial and temporal proximities are exploited by scanning the entire study area via overlapping space-time regions (STRs). Each STR represents a possible disease outbreak with a geometrical shape which is either a cylinder or rectangular prism. The base corresponds to the spatial dimension and the height corresponds to the temporal dimension. The dimensions of the STR are adjustable parameters. For instance, the maximum spatial dimension (e.g., the circular base of a cylindrical STR) can represent the maximum possible boundary of an outbreak, and the height of the STR could be the maximum allowable time to detect the outbreak. The dimensions of the STR are allowed to vary in order to detect outbreaks of varying sizes.

The initial proposition of STSS is based on the comparison of the disease rates inside and outside of a STR. If the disease rate inside the STR is significantly higher than outside the STR, then a possible disease outbreak is detected. However, this does not take into account the temporal variations (e.g., seasonal trends), which are inherent in epidemiological data. More recently, an expectation based approach was proposed to accommodate the temporal trend, where the observed value of an STR is compared with its expected value based upon historical data. Based on the statistical distribution that the data is assumed to follow, comparison is made via a likelihood ratio score function. If a STR has a likelihood ratio score bigger than 1, the STR is a potential disease outbreak. To reduce the false-alarm rate (reporting a disease outbreak where in reality there is no outbreak), the significance of the potential STRs is further tested via Monte Carlo simulation. If the STR is found to be significant at this stage, then a disease outbreak is recorded (Neill 2008)

STSS has the significant drawback that the entire study region has to be scanned, which is computationally intensive and limits the method’s scalability. Although previous research has shown that this problem can be tackled via efficient spatial-indexing methods. The assumption that a disease outbreak is a regular geometrical shape is also not realistic (e.g., disease might have spread via the river, thus affecting the people near the river bed) and remains as a limitation of the method. This problem might be tackled by generating irregularly shaped STRs.

60.5 Space-Time Visualization

Representing a phenomenon that evolves over space and time has emerged as a contentious issue within the GIS community. The contentious issue comes from the fact that most geographic phenomena change over time; for example, forest fires, storms, water contamination and also traffic congestion, but representing time on a map is still difficult. It is because GIS has its roots in mapping, which originally was designed to represent static phenomena, not dynamic process. Geographic visualization enhances traditional cartography by providing dynamic and interactive maps. Many new techniques on visualizing time on maps have been proposed. These techniques can be divided into three broad types: (static) 2D and 3D maps, and animation.

60.5.1 2D maps

There are various ways to represent time on static 2D maps, either as a single static map or multiple snap shots. Since all time steps are shown at the same time, the map-readers don't need to retain events temporarily in their minds thus preventing lapses of certain critical information. However, this technique can only present a few time steps at a time due to the limitation of the available map media (computer screen, paper, etc.).

This section will discuss some interesting static map techniques. The techniques are divided according to the type of data to be presented: geometric change of spatial objects (movement, size, shape, etc.), attribute change of spatial objects, and travel time.

60.5.1.1 Representing Geometric Change of Spatial Objects

Monmonier (1990) presents movement of spatial objects by drawing movement paths or pinpoints of objects on a 2D plane. Arrows are added to represent directions of movements. This technique is called a “dance map” since it is similar to a diagram of foot paths in a ballroom dance. Dance maps can display both discrete and continuous movement. When data are captured at fixed time intervals, a dance map can display the rate of movement (or rate of change) very well. Color or variety in sizes of objects can be added to the map, but the number of objects is limited by occlusion.

Another visualization technique presented by Monmonier (1990) is the *chess map* (map series). Each map contains a snap shot representing a time slice. A series of maps are laid out continuously in the manner of a chess board for users to compare events between time slices, allowing the comparison of many different time slices at a single sitting. The disadvantage of chess maps is that a large space is required to present multiple maps at the same time. In addition, the users must determine by themselves as to how the changes occurred, and at which time slices.

60.5.1.2 Representing Thematic Attribute Change of Spatial Objects

A *change map* shows changes or differences against a reference time period, as an absolute value or percentage, such as population increase every 10 years compared with 1990 (Monmonier 1990). The change map is good for representing quantitative attributes. Readers do not have to calculate the amount of change by themselves.

Another way to show the change is to add “small charts on maps” to visualize time series data on maps. The advantage of small charts on maps is that map readers are informed of the locations of the data on the maps as well as how their attributes change over time. However, when plotting many charts simultaneously, the base map can become overcrowded. Moreover, the charts can be easily overlapped when the data locations are very close to one another. An example of small charts on maps is given in Andrienko and Andrienko (2007).

Visual variables (colors, sizes, texture) can be applied to represent variation in attributes at different locations. The classic example of this technique is Minard's map showing Napoleon's doomed campaign to Moscow in 1869. Time was displayed as an axis on the map (parallel to the axis of the geographical position), and the number of remaining soldiers was shown by the thickness of the lines. Another good example is spatial treemaps (Slingsby et al. 2010) that represent traffic variables (traffic speed and traffic volume) of areas of London. Each grid cell on the map represents a borough. The level of brightness of each cell on spatial treemaps is used to represent the value of a traffic variable (speed and volume). Time is also mapped onto small cells within each area. This technique allows the visualization of a large number of time points, since it exploits every pixel on the map to represent data.

Rank Clock has recently been used to visualize the dynamics of city size changes (<http://www.bartlett.ucl.ac.uk/casa/pdf/paper152.pdf>), where the time is arranged as a clock, the thematic attribute (the size of the city) shows as a dot along the time line. By linking all the dots of a spatial unit over time, the trajectory of rank change is shown.

2D Space-time coloured pixels is widely used to study patterns of traffic congestion in space-time. It was used to display data from loop detectors. The space-time coloured pixels consists of two axes: a device position axis and a time axis. Each pixel represents the magnitude of traffic parameter, in colours, measured from a monitoring device at a particular time. Any anomaly of the detectors can also be shown easily by this approach.

60.5.1.3 Travel Time

The previous two subsections use time as a reference for other types of data (changes in geometry and in attributes of spatial objects). Here we pay special attention for travel time representation since time itself is the data to be represented and special techniques are developed for this purpose. There are two techniques that are used to present travel time on maps.

A *cartogram* is a map that distorts geographic space on maps to represent attributes of spatial data. For example, the tube map of London arrange all the tube lines in six zones in order to show the distance to the centre of London (Zone 1), which is not the exact physical (geometric) locations of the tube lines and stations. Using this technique, travel time on transportation networks can be represented using distance on a map, an example of which is the travel time tube map that distorts real geographic layout of tube lines in London in order to show travel time between stations.

The “isochrone” is another technique that is employed to represent travel time. Isochrones are similar to contour lines on a map, but an isochrone line connects points of equal travel time from a given origin (Brunsden et al. 2007). The isochrone is a great alternative to the cartogram as it does not distort the underlying map.

60.5.2 3D Visualization

60.5.2.1 3D Space-Time Cube

The 3D space-time cube (or, alternatively, space-time aquarium) was proposed by Hagerstrand (1970). A 3D space-time cube consists of two dimensions of geographic locations on a horizontal plane, and a time dimension in the vertical plane (or axis). The space-time cube is normally applied to represent trajectories of objects in 3D space-time dimension, or “space-time paths.” Trajectories are normally from GPS data nowadays, and they are represented as lines in the 3D space-time.

3D space-time cube has two main limitations. Firstly, the 3D display makes it difficult to refer space-time paths to geo-locations and time. Secondly, the space-time cube has difficulty in displaying large amounts of data. However, interactive techniques can be used to reduce cluttering when displaying a large amount of data. With interactive functions, users can decide which data to display and can zoom and rotate the cube on its axes. Data aggregation (such as generalized space-time path) can also improve visualization on 3D space-time cube.

60.5.2.2 3D Isosurface

An isosurface is a three-dimensional analog of an isoline. It is a surface that represents points of a constant value (e.g., pressure, temperature, velocity, density) within a volume of space. Isosurface has been employed in various applications such as medical imaging, fluid dynamics, astrophysics, chemistry and quantum mechanics. Isosurfaces are popularly used to visualize volumetric datasets, which consist of a 3D location with one scalar or vector attribute. The data sets are structured as (x,y,z,v) , where (x,y,z) are the spatial coordinates and v is an attribute. The 3D isosurface has also been applied to visualize incident data, which are structured as (x,y,t) , where (x,y) are two dimensional spatial coordinates and t is the time when the incident occurred (Brunsdon et al. 2007). Isosurfaces have great potential to show the development of space-time processes such as congestion on the traffic network.

60.5.2.3 3D Wall Map

The 3D wall map is a 2D road map with an additional time dimension to display change. Each layer represents the situation at a time. Cheng et al. (2010) employed the technique to represent travel delay during the morning peak in central London at four consecutive Mondays in October 2009. The layout of the link map represents the real geographical layout of the road network. The colours between layers represent the unit journey time (minutes per kilometre), with yellow and red colours showing the highly congested areas (travel time more than 5 min per kilometre) (Fig. 60.3).

60.5.3 Animated Maps

The first computer based animation map was created by Tobler in 1970 (Tobler 1970). He used 3D animated maps to display simulated urban growth data

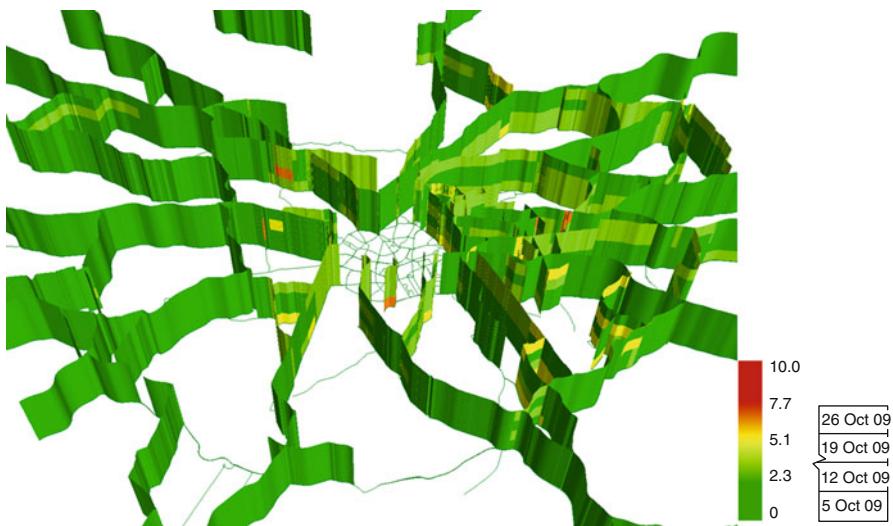


Fig. 60.3 Wall map of travel delay (mins/km) of outbound roads during the morning peak on 5,12,19,26 October 2009 (Cheng et al. 2011a)

in the Detroit region of the US. However, animated maps were not widely used for many years due to the difficulty in distributing and playing back large data files. However, with improvements in computing power and internet technology over the past three decades, animation maps have become a very active area of research and are now distributed widely on the internet. Weather maps and traffic maps are two of the many examples.

An animated map has two outstanding advantages. The first one is that an animation map can be used as an alternative to a static map. It can be employed to emphasize key attributes by using, for example, blinking symbols “to attract attention to a certain location on the map” (Kraak and Klomp 1995). The second advantage is that it provides additional visual variables called “dynamic variables” such as “duration,” “rate of change,” “order of change,” “frequency,” “display time,” and synchronization (MacEachren et al. 2004).

60.5.4 Visual Analytics: The Current Visualization Trend

Visual analytics is an outgrowth of the field of scientific and information visualization. It refers to “the science of analytical reasoning facilitated by interactive visual interfaces” (Thomas and Cook 2005). The emergence of visual analytics has been driven by the fact that we have no proper tools to leverage large amounts of data. Visual analytics is an iterative process that involves information gathering, data pre-processing, knowledge representation, and decision making. Normally, unknown data are visualized in order to give a basic view of that data, then users will use their perception (intuition) to gain further insights from the images

produced by visualization. The insights generated by this human perception are then transformed into knowledge. After users have gained certain knowledge, they can generate hypotheses that will be used to carry out further analysis using available data analysis and exploration techniques. The results from analytical process will be visualized for presentation, and further gain in knowledge.

Visual analytics is much more than simple visualization. It can rather be seen as an integral approach to combining visualization with human factors, and data analysis (Keim et al. 2008). Visual analytics is becoming more important to many disciplines including scientific research, business enterprise, and other areas that face problems of overwhelming avalanche of data. GIS, also, is now facing this massive data problem. The concept of visual analytics was introduced into GIS, namely “Geovisual Analytics”. Geovisual analytics has its specific focus on space and time; posing different specific research problems, and demands special approaches in solving generic research problems of Visual Analytics.

60.6 Conclusions

Since the concept of knowledge discovery from databases (KDD) was proposed in 1988, tremendous progress has been made in data mining and spatial data mining (Miller and Han 2009; Shekhar et al. 2011). STDM is only possible based upon the progress in those areas, along with GIS and geocomputation. This chapter introduces the fundamentals of STDM, which consists of space-time prediction, clustering and visualization.

As for space-time prediction, we have discussed the statistical (parametric) models, including families of STARIMA models, space-time geostatistical models, spatial panel data models, and space-time GWR. The challenge in statistical models lies in the non-stationary and non-linearity of space-time data. How to calibrate the spatio-temporal autocorrelations in the models is the bottleneck of statistical approaches. For low spatio/temporal resolution data, use of parametric statistical models may be preferable because of their explanatory power and interpretability. Due to their ability to deal with multi-dimensional nonlinear data machine learning methods are becoming more popular for large datasets. We have briefly introduced artificial neural networks (ANNs), support vector machines (SVMs), and other methods (Kernel-based approach and self-organized maps) for space-time analysis. However, the interpretability of machine learning is low, and a hybrid framework with both statistical and machine learning approaches might be helpful for this.

Space-time clustering can be used to extract meaningful patterns (clusters) in the data. It can also be used to detect outliers or emerging phenomena (epidemic outbreak or traffic congestion). Considering the spatial, temporal and thematic attributes seamlessly and simultaneously, and the dynamicity in the data is the most difficult challenge in spatio-temporal clustering. Spatio-temporal scan statistics (STSS) sheds lights on this aspect, though efforts are needed to improve computation efficiency and to reduce the false alarm rate.

Space-time visualization explores the patterns hidden in the large data sets by using advanced visualization and animation techniques. This includes conventional static 2D maps as well as newly developed 3D wall maps and isosurface, which shows the hotspots in space-time. Recently “Visual Analytics” and “Geovisual Analytics” have emerged as an iterative process (or tools) that involves information gathering, data pre-processing, knowledge representation/visualization, and decision making. Still, real-time visualization of dynamic processes is still very challenging due to large volume and high dimensions of the data. For examples, methods are needed to show the evolution and dissipation of crime or traffic congestion in space and time simultaneously.

However, the field of STDM is far from mature, and further research is needed in the following areas:

- a. New methods and theory are needed for mining crowd sources such as data contributed by citizens and volunteers. These are often extremely noisy, biased, and nonstationary. One example of such data is the trajectory data obtained from smart phones or other sensors. This area is relevant to the recent development of citizen sciences and VGI in particular.
- b. Theory and methods need to be developed to extract meaningful patterns from those individual sensors and put them under the framework of networks and network complexity such as transport and social-networks made up of those individual. Under network, the interaction and dynamic flows should be considered in mining spatio-temporal patterns. This aspect is relevant to complexity theory and network dynamics in particular.
- c. STDM for emergency and tipping points detection, leading to the generation of, actionable knowledge, i.e., finding the emergent patterns and tipping points of economic crises and disease epidemics. It is important to find outliers, but more important is finding the critical points before the system breaks down so that mitigating action can be taken to avoid the worst scenarios such as traffic congestion and epidemic transmission.
- d. Another challenge of STDM is how to calibrate, explain and validate the knowledge extracted. A good example of this is the calibration of spatial (or spatio-temporal) autocorrelation. Higher order spatial autocorrelation models have been developed, but the pitfalls have also been found (LeSage and Pace 2011). Nonstationarity and autocorrelation is fundamental to our observation (or our empirical test) of reality, it is hard to prove that the higher order autocorrelation comes from the first to the second, then to the third; or from the first to the third directly, which makes the explanation unconvincing. Furthermore, validation is difficult – so far Monte Carlo simulation is the main tool for simulation, which is also based upon a statistical distribution, which is hardly provable. This makes machine learning more promising in future STDM.
- e. Technically, grid computation and cloud computation allow data mining to be implemented at multiple computer sources. Even so, when the data volume is increased, the capacity of software and hardware is still limited. How to scale the algorithm to larger networks will always be a challenge for data mining given the

increase of data volume is far quicker than the improvement in the performance of data processors.

Please notice that the content of this chapter is mainly around spatial data in point, line and lattices, but not on image data, which is another broad area of research. Also, due to the limit of length, we do not include the progress on space-time simulation, which includes agent-based modeling (ABM) and cellular automata (CA). ABM has been used across many disciplines to demonstrate the impact of individual decisions and choices on the nature of a system (Gilbert 2007). Such examples include the individual behavior of birds in flocks, ants in colonies and people in crowds – all entities are acting independently yet contribute to a larger body. There is great potential within ABM to replicate and predict system changes over space and time. In (Manley et al. 2011), the agent-based simulation has demonstrated the link between individual choice and behavior in abnormal conditions with the formation and movement of urban road congestion. CA is a discrete model studied in computability theory, mathematics, physics, complexity science, theoretical biology and microstructure modeling. It consists of a regular grid of cells, each in one of a finite number of states, such as “On” and “Off.” It has been widely used in urban planning and landuse change modeling.

Acknowledgments This work is part of the STANDARD project – Spatio-Temporal Analysis of Network Data and Road Developments (standard.cege.ucl.ac.uk), supported by the UK Engineering and Physical Sciences Research Council (EP/G023212/1) and Transport for London (TfL).

References

- Anselin L (1988) Spatial econometrics: methods and models. Kluwer, Dordrecht
- Box G, Jenkins G (1970) Time series analysis: forecasting and control. Holden-Day, San Francisco
- Brunsdon C, Corcoran J, Higgs G (2007) Visualising space and time in crime patterns: a comparison of methods. *Comput Environ Urban Syst* 31(1):52–75
- Cheng T, Li Z (2006) A multi-scale approach for spatial-temporal outlier detection. *Trans GIS* 10(2):253–263
- Cheng T, Tanaksaranond G, Emmonds A, Sonoiki D (2010) Multi-scale visualization of inbound and outbound traffic delays in London. *Cartogr J* 47:323–329
- Cheng T, Haworth J, Wang J (2011a) Spatio-temporal autocorrelation of road network data. *J Geograph Syst*. <http://www.springerlink.com/content/4184v7072737621p/> Accessed 12 Oct 2011
- Cheng T, Wang J, Li X (2011b) A hybrid framework for space–time modeling of environmental data. *Geogr Anal* 43(2):188–210
- Elhorst JP (2003) Specification and estimation of spatial panel data models. *Int Reg Sci Rev* 26(3):244–268
- Fischer MM (2006) Spatial analysis and geocomputation. Springer, Berlin/ Heidelberg
- Gilbert N (2007) Agent-based models. Sage, London
- Griffith DA (2010) Modeling spatio-temporal relationships: retrospect and prospect. *J Geogr Syst* 12(2):111–123
- Hägerstrand T (1970) What about people in regional science? *Papers Reg Sci* 24(1):1–12
- Heuvelink GBM, Griffith DA (2010) Space-time geostatistics for geography: a case study of radiation monitoring across parts of germany. *Geogr Anal* 42(2):161–179

- Hsieh WW (2009) Machine learning methods in the environmental sciences: neural networks and Kernels, 1st edn. Cambridge University Press, Cambridge
- Huang B, Wu B, Barry M (2010) Geographically and temporally weighted regression for modeling spatio-temporal variation in house prices. *Int J Geogr Inf Sci* 24(3):383–401
- Kamarianakis Y, Prastacos P (2005) Space-time modeling of traffic flow. *Comput Geosci* 31(2):119–133
- Kanevski M, Timonin V, Pozdnukhov A (2009) Machine learning for spatial environmental data: theory, applications, and software. CRC Press, Boca Raton
- Keim D, Andrienko G, Fekete JD, Görg C, Kohlhammer J, Melançon G (2008) Visual analytics: definition, process, and challenges. *Inf Visual* 4950:154–175
- Kraak MJ, Klomp A (1995) A classification of cartographic animations: towards a tool for the design of dynamic maps in a gis environment. In: Proceedings of the seminar on teching animated cartography. Madrid, Spain, pp 29–35
- Kyriakidis PC, Journel AG (1999) Geostatistical space–time models: a review. *Math Geol* 31(6):651–684
- LeSage JP, Pace RK (2011) Pitfalls in higher order model extensions of basic spatial regression methodology. http://www.be.wvu.edu/econ_seminar/documents/11-12/lesage.pdf. Accessed on 15 Nov 2011
- MacEachren A, Gahegan M, Pike W, Brewer I, Cai G, Lengerich E, Hardisty F (2004) Geovisualization for knowledge construction and decision-support. *IEEE Comput Graph Appl* 24:13–17
- Manley E, Cheng T, Emmonds A (2011) Understanding route choice by using agent-based simulation. In: Proceedings of 11th international conference of geocomputation, London, 20–22 July 2011, pp 54–58
- Miller HJ, Han J (2009) Geographic data mining and knowledge discovery: an overview. In: Miller H, Han J (eds) *Geographic data mining and knowledge discovery*, 2nd edn. Taylor and Francis, Boca Raton
- Monmonier M (1990) Strategies for the visualization of geographic time-series data. *Cartographica* 27(1):30–45
- Neill DB (2008) Expectation-based scan statistics for monitoring spatial time series data. *Int J Forecast* 25(3):498–517
- Pfeifer PE, Deutsch SJ (1980) A three-stage iterative procedure for space-time modeling. *Technometrics* 22(1):35–47
- Pozdnoukhov A, Matasci G, Kanevski M, Purves RS (2011) Spatio-temporal avalanche forecasting with support vector machines. *Nat Hazards Earth Syst Sci* 11:367–382
- Rey SJ, Janikas MV (2010) STARS: Space-time analysis of regional systems. In: Fischer MM, Getis A (eds) *Handbook of applied spatial analysis: software tools, methods and applications*. Springer, Berlin/Heidelberg, pp 91–112
- Shekhar S, Evans MR, Kang JM, Mohan P (2011) Identifying patterns in spatial information: a survey of methods. *Wiley Interdiscip Rev Data Min Knowl Discov* 1(3):193–214
- Slingsby A, Wood J, Dykes J (2010) Treemap cartography for showing spatial and temporal traffic patterns. *J Maps* 2010:135–146
- Thomas JJ, Cook KA (2005) Illuminating the path: the research and development agenda for visual analytics. IEEE, Los Alamitos
- Tobler W (1970) A computer movie simulating urban growth in the detroit region. *Econ Geogr* 46:234–240
- Vapnik V (1999) The nature of statistical learning theory, 2nd edn. Springer, London

Chris Brunsdon

Contents

61.1	Introduction	1195
61.2	Kinds of Spatial Data	1197
61.3	Bayesian Approaches for Point Data	1198
61.4	A Roughness-Based Prior	1202
61.5	Bayesian Approaches for Point-Based Measurement Data	1206
61.6	Region-Based Measurement Data	1212
61.7	Conclusions	1213
	References	1215

Abstract

This chapter outlines the key ideas of Bayesian spatial data analysis, together with some practical examples. An introduction to the general ideas of Bayesian inference is given, and in particular the key rôle of MCMC approaches is emphasized. Following this, techniques are discussed for three key types of spatial data: *point data*, *point-based measurement data*, and *area data*. For each of these, examples of appropriate kinds of spatial data are considered and examples of their use are also provided. The chapter concludes with a discussion of the advantages that Bayesian spatial analysis has to offer as well as considering some of the challenges that this relatively new approach is faced with.

C. Brunsdon

School of Environmental Sciences, University of Liverpool, Liverpool, UK

e-mail: Christopher.Brunsdon@liverpool.ac.uk

61.1 Introduction

Bayesian analysis has seen an enormous increase in popularity over recent years. There are a number of possible reasons for this. Firstly, a framework in which prior beliefs may be incorporated can offer certain advantages. Particularly, one can examine a new study in the light of findings of previous studies – something which cannot be done in a classical framework. Secondly, inferences drawn are based on posterior distributions for parameters of interest. Some find this a more intuitive basis for inference than classical significance tests and confidence intervals. Indeed, one author considering the teaching of elementary statistics observed that some of his students misinterpreted classical inference in a way that coincided with Bayesian inference (Berry 1997). Thirdly, a probability distribution for a parameter contains more information than a classical confidence interval. For example, a bimodal posterior or a highly skewed posterior would offer a more subtle interpretation of the outcome of a study than a simple interval or point estimate.

To recall the basic ideas of Bayesian inference, an individual supplies information *prior beliefs* and some unobservable parameter (or set of parameters) θ in the form of a *probability distribution* or *probability density function* $f(\theta)$. If the observable data \mathbf{x} has a likelihood function $L(\mathbf{x}|\theta)$ then Bayes theorem can be used to obtain the relationship

$$f(\theta|\mathbf{x}) \propto f(\theta)L(\mathbf{x}|\theta) \quad (61.1)$$

thus giving an expression for the probability distribution of the unobservable θ given the observable data \mathbf{x} . This is a very different framework to the classical or *frequentist* approach to statistical inference above θ . Although the latter makes use of $L(\mathbf{x}|\theta)$, θ itself is treated as a deterministic, but unobservable, quantity. As such, $f(\theta)$ and $f(\theta|\mathbf{x})$ are not considered as relevant concepts, and instead, hypotheses as to whether statements about θ are true are used as the basis for inference. The need to supply $f(\theta)$ in Bayesian inference is a notable qualitative distinction between the two approaches, as it requires the analyst to supply a subjective set of beliefs about θ as part of the analysis process – although these beliefs could represent a state of impartiality (e.g., by supplying a uniform distribution for θ). In this situation the prior distribution is often referred to as a *noninformative prior*.

One practical difficulty with this approach is that Eq. (61.1) only defines the posterior distribution up to a constant of proportionality. To normalize the distribution (so that the integral over all θ values is one), the equation

$$f(\theta|\mathbf{x}) = \frac{f(\theta)L(\mathbf{x}|\theta)}{\int f(\theta)L(\mathbf{x}|\theta)d\theta} \quad (61.2)$$

is used.

However, the integral in the denominator of Eq. (61.2) is not always analytically soluble, which in the past has led to some difficulties with carrying out Bayesian

analysis in practice. Fortunately, recent advances in computational techniques have addressed this issue. In particular, the advent of Markov Chain Monte Carlo (MCMC) simulation-based approaches has allowed Bayesian approaches to be applied in a wide variety of situations where analytical results are intractable. The ideas of Bayesian inference, in general, and of MCMC methods are covered in ▶ Sect. 9, “[Spatial Econometrics](#)” in this major reference work. The aim of this chapter is to focus on how these ideas can be applied to the analysis of *spatial* data. In many ways there is nothing special about Bayesian techniques for spatial data – essentially the principles underlying the inferential process (stemming from Bayes’ theorem) are the same as those used for any kind of data. However, although Bayesian analysis of spatial data may have the same overarching framework, there are distinct characteristics of the kinds of *model* to which it is applied and also to the kinds of *data structure* one is likely to encounter. Thus, in this chapter, models in which the random components are spatially correlated and the ways in which Bayesian inference is made about spatial characteristics of the modeled processes will be considered. Firstly, the kinds of spatial data with which Bayesian methods are commonly used will be outlined. Next, examples of how Bayesian methods are applied for each type of data will be given. Finally some practical considerations will be considered.

61.2 Kinds of Spatial Data

Elsewhere in this book, a comprehensive list of spatial data types is provided. However, not all of these data types may be analyzed in a Bayesian framework using techniques that are well established at the time of writing. In particular, there are few Bayesian methods that may be applied to arc- or line-based data. Following the typology of Fischer and Wang (2011), some kinds of data for which well-established Bayesian approaches exist are now listed:

- *Point data*: These are data consisting of a set two or three dimensional point coordinates in Euclidean space. The points themselves are considered as random and typically interest lies in modeling a stochastic *spatial point process* that could have generated the data.
- *Point measurement data*: These are data consisting of a set of spatial points, as before, but here each point has an attached *attribute*. Typically the attribute is some kind of measurement, such as a temperature taken at that location or the price of a house sold at that location. Here, in general it is only the measured attribute that is assumed to be a random component of the model, the locations being treated as fixed, controlled values – effectively part of the design of the data collection procedure. Typically, the spatial component of models for this kind of data arises from an assumption that correlations between the attributes is in some way dependent on the relative locations of the points.
- *Field data*: These are data that relate to variables which are conceptually continuous (the field view) and whose observations have been sampled at an

pre-defined and fixed set of point locations. Arguably, those have a great deal in common with point measurement data – although those are defined *only* at a fixed set of points, while field data relate to a *sample* of point measurements of a mapping from *all* points in space to an attribute value. In this case, a measurement (such as temperature) could have been taken at *any* point in space.

- *Area data:* These are data consisting of a set of spatial regions – typically represented as polygons. In most studies, the polygons provide a partition of a geographical study area, that is, their union completely covers the study area, and no pair of polygons intersects, except in some cases on their boundaries. In the latter situation, regions with boundaries touching are said to be *adjacent*. As with point-based measurement data, an attribute is associated with each entity – the entities now being regions instead of points. Also in common with point-based measurement data, the regions are considered to be fixed (not arising from a random process), and the only component assumed random in models of these data is the attribute. In this case, the spatial aspect of the model is achieved by relating the correlation structure of the random attributes to the relative position of regions – in particular, they often make use of the adjacency or otherwise of each pair of regions in the data set.
- *Spatial interaction data:* These data (also termed origin-destination flow or link data) consist of measurements or counts, each of which is associated with a pair of point locations, or pair of areas. For example, travel to work data, listing the number of people traveling from a given origin (home) zone to a given destination (workplace) zone, fall into this category.

In this chapter, attention will be focused on point data, point measurement data, and area data – as in Cressie (1993), although the technique suggested for point measurement data may also be applied to field data.

61.3 Bayesian Approaches for Point Data

A key model for point data is the *spatial Poisson process*. In such a process, location of points occurs independently of one another, with the *intensity* of occurrences in location $s = (s_1, s_2)$ being given by $\Lambda(s)$. For any area A within the region of study, the number of occurrences has a Poisson distribution with mean

$$\int_{s \in A} \Lambda(s) ds \quad (61.3)$$

In such models, it is informative to estimate $\Lambda(s)$ since mapping this function allows regions of high intensity to be mapped, and related features, such as areas of high variability in intensity, or locations of peak intensity can be investigated. One very simple method of estimation is to use a *pixel-* or *regular lattice*-based approach. Suppose the study area is partitioned into a number of small, identical

tesselating polygons (typically squares or regular hexagons), Δ_k for $k \in \{1, \dots, n\}$. If λ_k is the mean value of $\Lambda(\mathbf{s})$ for $\mathbf{s} \in \Delta_k$, then

$$\lambda_k = \frac{1}{A} \int_{\mathbf{s} \in \Delta_k} \Lambda(\mathbf{s}) d\mathbf{s} \quad (61.4)$$

where A is the area of each Δ_k . If the value of A is reasonably low, so that each Δ_k occupies a very small proportion of the study area, it is reasonable to assume that $\Lambda(\mathbf{s})$ does not vary much within each Δ_k , so that the lattice based approximation

$$\Lambda(\mathbf{s}) = \lambda_k \text{ for } \mathbf{s} \in \Delta_k \quad (61.5)$$

may be used. Now, if c_k is the number of points occurring in Δ_k , then c_k has a Poisson distribution with mean $A\lambda_k$. For simplicity, assume for now that spatial units are chosen such that $A = 1$ – this simplifies equations, without any loss of generality. With this assumption in place, c_k has a Poisson distribution with mean λ_k . At this stage, a Bayesian approach may be applied to estimate λ_k , since the Poisson distribution of c_k gives

$$\Pr(c_k | \lambda_k) = \exp(-\lambda_k) \lambda_k^{c_k} \quad (61.6)$$

and so if the prior probability density function for λ_k is $f(\lambda_k)$, then the posterior distribution $f(\lambda_k | c_k)$ is related to the prior and the likelihood function by

$$f(\lambda_k | c_k) \propto f(\lambda_k) \exp(-\lambda_k) \lambda_k^{c_k} \quad (61.7)$$

In particular, if the prior for λ_k is a gamma distribution, proportional to $\lambda_k^{\alpha-1} \exp(-\beta\lambda_k)$ for $\lambda_k > 0$, and zero otherwise, and where the constant of proportionality is independent of λ_k , then we have

$$f(\lambda_k | c_k) \propto \exp(-(1 + \beta)\lambda_k) \lambda_k^{c_k + \alpha - 1} \text{ if } \lambda_k > 0, \text{ and zero otherwise} \quad (61.8)$$

which is also a gamma distribution for λ_k , with updated parameters $\beta' = \beta + 1$ and $\alpha' = \alpha + c_k$. In particular, in the case where $\alpha = \beta = 0$, we have $f(\lambda_k) \propto \lambda_k^{-1}$. This is an example of an improper prior distribution – although it is not a well-defined probability function itself, the corresponding posterior distribution is well defined, being a gamma distribution with parameters $\alpha = c_k$ and $\beta = 1$. In this case, the expected value of the posterior distribution for λ_k is just c_k . This value may be used to provide a point estimate of λ_k for each area Δ_k .

As an example, consider the inventory data of the Zurichberg Forest, Switzerland (see Mandallaz (2008) for details), which lists the locations (and types) of trees in the forest. These data are provided with the kind authorization of the Forest Service of the Canton of Zurich. Figure 61.1 shows the raw tree

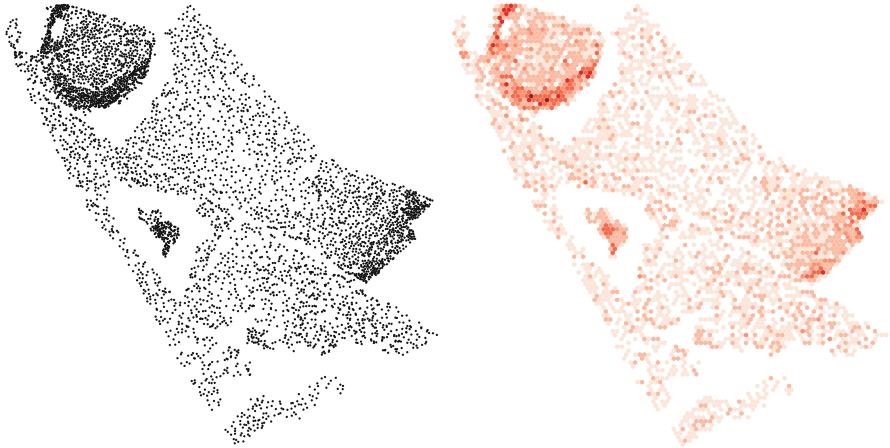


Fig. 61.1 Raw trees data from the Zurichberg Forest (RHS) and estimates of $\Lambda(s)$ based on a hexagonal grid and noninformative priors for the finite elements λ_k (LHS)

location data and the estimated values of $\Lambda(s)$ using the above approach, with the Δ_k zones being elements of a hexagonal tesselating grid.

One characteristic of this kind of estimate is that although it is only a minor generalization of the raw data (it simply estimates the intensity in Δ_k as being proportional to the observed count of events in that area, c_k), the underlying Bayesian theory also provides a posterior distribution, so that one could, for example, estimate other features of the posterior distribution for λ_k such as the upper 95th percentile for each λ_k , or the summed intensity over some arbitrary region of the forest. In either case, posterior distributional features could be estimated (possibly via simulation) in addition to point estimates. As an example, consider the problem of estimating the average intensity over the entire forest. Assuming there are n Δ_k elements covering the forest, this quantity is

$$\frac{1}{nA} \int_{s \in \mathcal{F}} \Lambda(s) ds \text{ where } \mathcal{F} = \bigcup_{k=1, \dots, n} \Delta_k \quad (61.9)$$

If length units are chosen such that $A = 1$ as before, this is equal to

$$\frac{1}{n} \sum_{k=1, \dots, n} \lambda_k \quad (61.10)$$

which is the mean of all of the λ_k values. For brevity, this will now be denoted as $\bar{\lambda}$. If each λ_k has a posterior distribution as set out in Eq. (61.8), then it can be seen that

$$f(\bar{\lambda} | \{c_k, k = i, \dots, n\}) \propto \bar{\lambda}^{n\bar{c}-1} \exp(-n^{-1}\bar{\lambda}) \text{ if } \bar{\lambda} > 0, \text{ and zero otherwise} \quad (61.11)$$

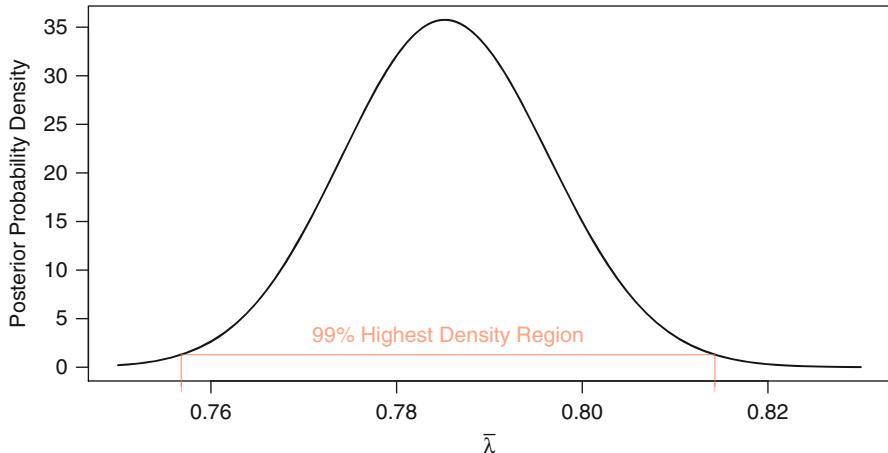


Fig. 61.2 Posterior probability distribution of mean intensity of trees in forest. 99 % HDR is also indicated

where \bar{c} is the arithmetic mean of the c_k 's. Thus the posterior distribution of $\bar{\lambda}$ is also gamma, with parameters $n\bar{c}$ and n^{-1} . Thus, the posterior distribution of $\bar{\lambda}$ is shown in Fig. 61.2.

To obtain an interval estimate of $\bar{\lambda}$, the $a\%$ *Highest Density Region* (HDR) (Hyndman 1996) can be found. This is the region in the set of possible values of $\bar{\lambda}$ such that the posterior probability that $\bar{\lambda}$ lies in the region is $a/100$, and the posterior probability density of any point within the region is higher than any point outside of the region. For the Zurichberg Forest data, the 99 % HDR is the interval (0.757, 0.814). This is indicated on Fig. 61.2.

The above example used a prior probability density distribution that assumed independence between the individual discretized intensity levels λ_k . This assumption is manifested in certain characteristics of the intensity estimates. In Fig. 61.1 the estimates show quite a large amount of spatial “roughness,” that is, there are a number of Δ_k zones whose mean intensity estimates λ_k are very different in value from their neighbors. However, there may be reason to expect that in fact these values should vary smoothly. If this is the case, Bayesian inference can be a useful tool since such expectations of smoothness can be expressed via the prior distribution.

An important issue is now to define roughness. In terms of any function $G(s)$, rather than the discrete approximation, one possible measurement of roughness is the expression

$$\mathcal{R}(G) = \int_{s \in \mathcal{F}} \left(\frac{\partial^2 G}{\partial s_1^2} + \frac{\partial^2 G}{\partial s_2^2} \right)^2 ds \quad (61.12)$$

where $s = (s_1, s_2)$. Note that the two partial second derivatives in Eq. (61.12) measure rate of change of slope and that if both of these have high positive or

high negative values, this indicates sharp maxima or minima – hence, squaring the sum of these and integrating over the study area is a measure of the propensity of $G(\mathbf{s})$ to have sharp peaks and pits and is therefore a plausible measure of roughness. Note also that $\mathcal{R}(G) = 0$ if and only if

$$\frac{\partial^2 G}{\partial s_1^2} + \frac{\partial^2 G}{\partial s_2^2} = 0 \quad (61.13)$$

that is, if $G(\mathbf{s})$ is a solution of Laplace's equation – frequently referred to as a *harmonic function*. Harmonic functions exhibit the *mean value property*

$$G(\mathbf{s}) = \frac{1}{2\pi r} \oint_{\mathbf{t} \in P(\mathbf{s}, \mathbf{r})} G(\mathbf{t}) d\mathbf{t} \quad (61.14)$$

if G is a harmonic function, and $P(\mathbf{s}, \mathbf{r})$ denotes a circular path of radius r centered on the location \mathbf{s} , provided that all of $P(\mathbf{s}, \mathbf{r})$ lies strictly within the region in which $G(\mathbf{s})$ is defined. In less mathematical terms, this states that the value of a harmonic function at a given point is equal to the mean value of that function taken over a circular path centered on that point. Again, this can be thought of as a condition of smoothness. Returning to Eq. (61.12), $\mathcal{R}(G)$ can be thought of as a measurement of the discrepancy between G and a harmonic function, and, that considered from this viewpoint, this provides an alternative interpretation of this quantity as a measure of roughness.

61.4 A Roughness-Based Prior

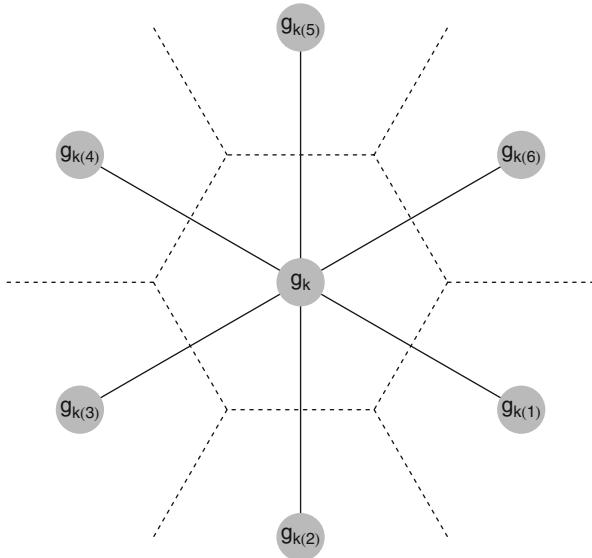
In the previous section, a measurement of roughness was proposed. In this section, this measure will be used to construct a Bayesian prior encapsulating subjective expectations of smoothness in $\Lambda(\mathbf{s})$ which will be combined with observed point data to obtain a posterior distribution for intensity. To do this, the roughness of some function G that is related to Λ may be used to construct a prior probability density function for the value of $\mathcal{R}(G)$ with an exponential form suggested:

$$f(\mathcal{R}(\Lambda)) \propto \kappa \exp(-\kappa \mathcal{R}(G)) \quad (61.15)$$

defined for positive $\mathcal{R}(G)$. The mean of this prior distribution is κ^{-1} so that low values of κ suggest that a high degree of roughness is expected.

Here, G will be related to Λ by $G(\mathbf{s}) = \sqrt{\Lambda(\mathbf{s}) + 3/8}$ – this is chosen since the transformation gives a distribution close to normal and has a variance-stabilizing effect, when the counts in the zones Δ_k have a Poisson distribution (Anscombe 1948; Mäkitalo and Foi 2011). In practice, as in the previous example, a finite

Fig. 61.3 Diagram showing notation for numerical values of neighbors of zone Δ_k (the central hexagonal zone)



element approach is used to estimate $G(\mathbf{s})$. In particular, it can be shown that $\mathcal{R}(G)$ can be approximated by

$$R = \text{const} \sum_{k=1,n} \left(g_k - \frac{1}{6} \sum_{i=1,6} g_{k(i)} \right)^2 \quad (61.16)$$

where $g_{k(i)} = \sqrt{\lambda_{k(i)}} + 3/8$ indicates the mean value G associated with each of the neighboring tesselating zones to Δ_k as set out in Fig. 61.3. Provided δ is reasonably small, these mean values $\{g_k, g_{k(1)}, g_{k(2)}, g_{k(3)}, g_{k(4)}, g_{k(5)}, g_{k(6)}\}$ are close to the sampled values of G at points annotated on the figure.

To see how this estimate is derived, firstly assume that each of the lines emanating from the central g_k has a length δ . To estimate $\partial^2 G / \partial s_1^2$ and $\partial^2 G / \partial s_2^2$ at the point (s_1, s_2) , a quadratic surface of the form

$$Q(s_1 + t_1, s_2 + t_2) \approx \gamma_0 + \gamma_1 t_1 + \gamma_2 t_2 + \gamma_{11} t_1^2 + \gamma_{12} t_1 t_2 + \gamma_{22} t_2^2 \quad (61.17)$$

is fitted to the values g_k and $g_{k(1)}, \dots, g_{k(6)}$ at the (t_1, t_2) locations corresponding to the central point and six hexagonally arranged lines in Fig. 61.3, respectively, using least squares approximation. The locations of the seven points are

$$\left\{ (0,0), \left(\frac{\sqrt{3}}{2}\delta, -\frac{1}{2}\delta\right), (0, -\delta), \left(-\frac{\sqrt{3}}{2}\delta, -\frac{1}{2}\delta\right), \left(-\frac{\sqrt{3}}{2}\delta, \frac{1}{2}\delta\right), (0, \delta), \left(\frac{\sqrt{3}}{2}\delta, \frac{1}{2}\delta\right) \right\} \quad (61.18)$$

and the corresponding Q values are $\{g_k, g_{k(1)}, g_{k(2)}, g_{k(3)}, g_{k(4)}, g_{k(5)}, g_{k(6)}\}$. Using this approximation and applying second partial derivatives, we obtain the approximations $\partial^2 G / \partial s_1^2 \approx \gamma_{11}$ and $\partial^2 G / \partial s_2^2 \approx \gamma_{22}$, so that

$$\frac{\partial^2 G}{\partial s_1^2} + \frac{\partial^2 G}{\partial s_2^2} \approx \gamma_{11} + \gamma_{22} \quad (61.19)$$

It may be checked that, when the values for the point locations and intensities are substituted into the formula for the least squares fitting, we have

$$\gamma_{11} + \gamma_{22} = -2 \left(\frac{1}{6} \sum_{i=1,6} g_{k(i)} - g_k \right) \quad (61.20)$$

and note also, from the Taylor expansion of $G(\mathbf{s} + \mathbf{t})$, that this approximation tends asymptotically to the true value as δ tends to zero. Finally, it is interesting to note that this result can be used to derive a discrete version of the mean value property, with the circular path around a point being approximated by the centroids of the neighboring hexagonal lattice elements.

Now, returning to the definition of roughness in Eq. (61.12), the approximation in Eq. (61.16) may be obtained. Entering this expression for R (as an approximation for \mathcal{R}) into Eq. (61.15) yields a roughness-penalty-based prior distribution for each g_k :

$$\Pr(g_k | g_{k(1)}, \dots, g_{k(6)}) \propto \kappa \exp \left(-\kappa \sum_{k=1,n} \left(g_k - \frac{1}{6} \sum_{i=1,6} g_{k(i)} \right)^2 \right) \quad (61.21)$$

It may be noted that this distribution takes the form of a *intrinsic conditional autoregressive* (ICAR) model (Besag 1974; Besag and Kooperberg 1995) – with precision parameter κ . This is an improper prior, as it does not have a well-defined multivariate distribution for the vector \mathbf{g} ; however, in conjunction with certain likelihood functions (e.g., multivariate normal), the posterior probability density is well defined. Thus, a prior distribution for the g_k values is constructed, and therefore, one for the λ_k values can be derived. However, this prior distribution requires the parameter κ to be provided. One possibility – if the analyst has a clear idea of the degree of roughness to expect – is to specify a particular value in advance. However, in many situations one cannot realistically do this, and another approach – demonstrated here – is to specify a *hyperprior* for the quantity. In this case, Eq. (61.21) is assumed to be conditioned on κ as well, with the prior for κ being an improper prior equal to a small constant.

Having constructed this prior distribution for the parameters, it is now necessary to consider the posterior distribution. Since we are working with the *transformed* parameters $g_k = \sqrt{\lambda_k + 3/8}$, the counts in each zone Δ_k will also be transformed using the same function, that is, c_k , the count of trees in each Δ_k , will be

transformed to $c'_k = \sqrt{c_k + 3/8}$. Anscombe (1948) suggests that Poisson counts from a distribution with mean λ_k transformed in this way have an approximately normal distribution with mean g_k and variance 1/4. Making use of this approximation, a degree of algebraic manipulation shows that the conditional posterior distributions for the g_k 's and κ are then given by

$$\begin{aligned} \Pr(g_k|g_{k(1)}, \dots, g_{k(6)}, \kappa) &\propto N\left(\frac{1}{24} \sum_{i=1,\dots,6} g_{k(i)} + \frac{c'_k}{\kappa}, \frac{1}{\sqrt{\kappa+4}}\right) \\ \Pr(\kappa|g_1, \dots, g_n) &\propto \text{Gamma}\left(n, \sum_{k=1,\dots,n} \left(g_k - \frac{1}{6} \sum_{i=1,6} g_{k(i)}\right)^2\right) \end{aligned} \quad (61.22)$$

where $N(\cdot)$ and gamma denote the normal and gamma distributions with the standard parameterizations. This specification of posterior probabilities lends itself to a Markov chain Monte Carlo (MCMC) approach. Rather than estimate the parameters of interest analytically, this approach draws simulated samples from the posterior distribution of the parameters of interest. Essentially, if estimates of all parameters except one are provided, the remaining one is then drawn from one of the conditional distributions shown in Eq. (61.22). Cycling through the parameter space, provided each parameter is drawn from the correct conditional distribution, a draw of all parameters from the full posterior distribution is obtained. If a large number of such multivariate draws are provided, then the empirical distribution may be used to estimate features of this posterior probability distribution. In this case, an MCMC approach is used to estimate and map the posterior mean of each λ_k , which is achieved by transforming the simulated distribution of the g_k 's, via the inverse Anscombe transform $\lambda_k = g_k^2 - 3/8$. The results are shown in Fig. 61.4.

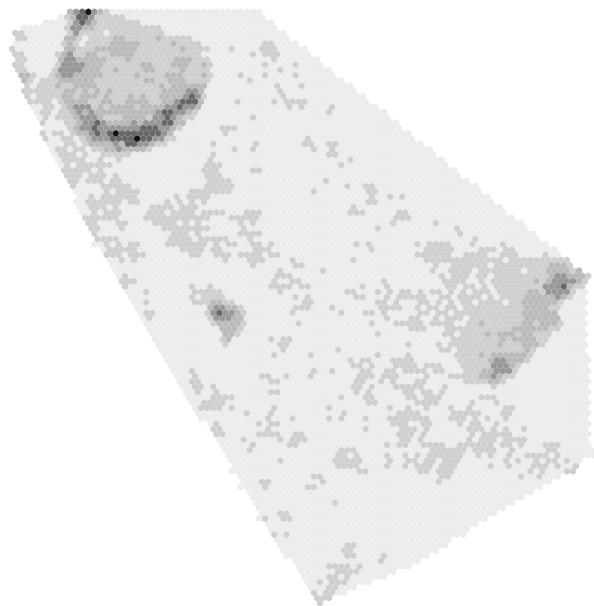
A further advantage of the MCMC approach is that estimation of quantities related to the λ_k values may be estimated in a very natural way, by simply computing these quantities for the simulated posterior λ_k value and viewing the resulting distribution of the computed quantities. For example, a quantity of interest may be the *slope* of $\Lambda(\mathbf{s})$. This is defined to be the magnitude of the vector $\nabla\Lambda(\mathbf{s})$. As before, a discrete approximation will be used – and this may be obtained from the coefficients of the quadratic expression Q in Eq. (61.17). In fact

$$\begin{aligned} \frac{\partial\Lambda}{\partial s_1} &\approx \gamma_1 \\ \frac{\partial\Lambda}{\partial s_2} &\approx \gamma_2 \end{aligned} \quad (61.23)$$

so that

$$|\nabla\Lambda(\mathbf{s})| = \sqrt{\left(\frac{\partial\Lambda}{\partial s_1}\right)^2 + \left(\frac{\partial\Lambda}{\partial s_2}\right)^2} \approx \sqrt{\gamma_1^2 + \gamma_2^2} \quad (61.24)$$

Fig. 61.4 Posterior mean estimates of trees in forest, based on implementing the MCMC algorithm set out in Eq. (61.22)



It may be checked that by using a least squares estimation of Q ,

$$\begin{bmatrix} \gamma_1 \\ \gamma_2 \end{bmatrix} = \frac{1}{3} \begin{bmatrix} \frac{\sqrt{3}}{2} \lambda_2 + \frac{\sqrt{3}}{2} \lambda_3 - \frac{\sqrt{3}}{2} \lambda_5 - \frac{\sqrt{3}}{2} \lambda_6 \\ \lambda_1 + \frac{1}{2} \lambda_2 - \frac{1}{2} \lambda_3 - \lambda_4 - \frac{1}{2} \lambda_5 + \frac{1}{2} \lambda_6 \end{bmatrix} \quad (61.25)$$

which may be used to obtain an estimate of the slope. Applying this to the MCMC simulations obtained earlier yields the map in Fig. 61.5. This shows the regions in which the tree density undergoes most change. As a final approach, one could designate any area where the slope exceeds 1.5 to be a *transition zone*. The idea of this is to identify boundary regions around the forest, for example, these could equate to *ecotones* (Holland and Risser 1991; Allen and Breshears 1998). Again, a function is applied to the simulated λ_k values to obtain a binary variable T_k , which takes the value one if λ_k exceeds 1.5, and zero otherwise. By counting the number of times (in 1,000 simulations) that T_k is equal to one, an estimate of the posterior probability that each zone Δ_k is part of a transitional area may be estimated. In addition, the posterior standard deviations are also illustrated (Fig. 61.6).

61.5 Bayesian Approaches for Point-Based Measurement Data

In this section, analysis of a set of data points will be considered, with each point having an attached attribute. In this case, the spatial point locations will not be treated as random, for example, they may be points where a set of measurements were taken, such as soil conductivity or rainfall levels or the rental or sale prices of

Fig. 61.5 Posterior mean estimates of slope of $\Lambda(\mathbf{s})$

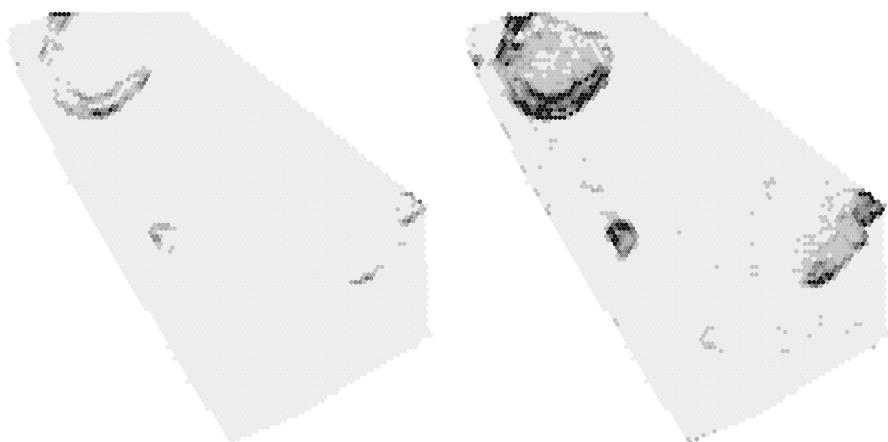


Fig. 61.6 Posterior mean estimates of T_k (left) and associated standard deviations (right – see text)

houses at a given set of addresses. The attributes will, however, be treated as a random spatial process. Typically, this will be done by specifying a *variogram* or *correlation function* specifying the degree of correlation between the attributes associated with a pair of points. For a *stationary process*, this correlation will depend only on the distance between the point pairs – so that measurements associated with any pair of points separated by a given distance will have the

same correlation, regardless of their absolute spatial location. Analyzing data using a model of this kind, particularly with a view to interpolating measured values, is referred to as *kriging* – based on the early work in this area developed by Matheron; see (Matheron 1970, 1973) for example.

The underlying idea here is that the attribute values are observations from a *field* defined as a real-valued random function over a space S . The randomness can occur in two ways. Firstly, the function itself is modeled as a Gaussian random process, with a value $E(z) = m(\mathbf{s})$, where $E(z)$ is the expected value of the attribute value z and $m(\cdot)$ is a function of location $\mathbf{s} \in S$. For *ordinary kriging* this function is just a constant value, μ . For a fuller discussion, see Cressie (1993). The covariance between two values at locations \mathbf{s}_1 and \mathbf{s}_2 is given by $\sigma^2 \rho(\mathbf{s}_1, \mathbf{s}_2)$, where $\rho(\mathbf{s}_1, \mathbf{s}_2)$ is the correlation between the z -values at \mathbf{s}_1 and \mathbf{s}_2 – and if the process depends only on the distance between the points, as stated above, then the $\rho(\cdot)$ function can be written in the form $\rho(|\mathbf{s}_1 - \mathbf{s}_2|)$, or $\rho(d)$ if $d = |\mathbf{s}_1 - \mathbf{s}_2|$. The relationship is often expressed in the form of a *variogram*:

$$\gamma(d) = \sigma^2(1 - \rho(d)) \quad (61.26)$$

A number of possible functional forms for γ are frequently used. Some following examples include

Exponential	$\gamma(d) = \sigma^2 \left[1 - \exp\left(-\frac{d}{h}\right) \right]$	
Spherical	$\gamma(d) = \sigma^2 \left(\frac{3d}{2h} - \frac{d^3}{2h^3} \right)$ if $d < h$; σ^2 otherwise	(61.27)
Gaussian	$\gamma(d) = \sigma^2 \left[1 - \exp\left(-\frac{d^2}{2h^2}\right) \right]$	

In each of these cases, the parameter h controls the amount of correlation between attribute values at locations separated by a given distance. Although the exact interpretation differs for each of the functions, it is generally the case that larger values of h suggest correlation persists at larger distances. Note that the list above is not exhaustive, but also that arbitrary specification of these functions is not possible – in general, functions must be chosen such that for any set of points in S , the covariance matrix implied by the function must be positive definite.

When analyzing this kind of data, there are generally two key issues requiring investigation. Firstly, the calibration of the variogram function, in the particular forms above, implies the estimation of the parameters h and σ^2 . The second issue is the use of this model for interpolation. Although the Gaussian process is sampled at a finite number of locations, it is often useful to estimate values of $E(z)$ at other locations. In particular, given that there is a vector of observed attribute values \mathbf{x} , interpolation at a new point \mathbf{s} can be thought of as estimating the conditional mean of z at this location, given \mathbf{x} – written as $E(z|\mathbf{x})$. In both cases, although the primary aim may be the investigation of parameters, being able to make statements about

the precision or accuracy of these statements is of importance. The original approach to kriging calibrates the variogram using point estimates for the parameters, and then “plugs them in” to the expression for $E(z|\mathbf{x})$. However, more recently, attention has focused on approaches that allow for uncertainty in the variogram parameter estimates. In terms of interpolation, these provide a more realistic picture of kriging estimates – although the original approaches provide expressions for variance in $E(z|\mathbf{x})$, these are conditioned on the parameter values in the variogram. In reality these parameters are generally estimated and subject to uncertainty due to the sampling process – and this in turn adds uncertainty in the estimates for $E(z|\mathbf{x})$.

A Bayesian approach – such as that given by Diggle et al. (1998) – is one way of addressing this issue. Prior probability distributions are supplied for the variogram parameters, and on the basis of observations, a posterior distribution function is derived. In the reference above, as in the previous section, an approach using Markov chain Monte Carlo simulations is used.

The technique is illustrated here with a set of house price data obtained from Nestoria (<http://www.nestoria.co.uk>) – a web site providing listings of houses currently for sale. Here a set of 260 semidetached three bedroom houses were downloaded for the Liverpool area of the United Kingdom. This set contains all three bedrooomed semidetached houses listed on 28th December 2011, excluding those whose price was not published in the listings. (A very small number of houses are listed as “price on application”.) Prices were scaled to units of thousands of pounds, to avoid rounding errors in calculations; thus, an asking price of \$199,950 would be recorded as \$199.95 k. Here, house locations are recorded as latitude and longitude and then transformed to the OS National Grid projection coordinates in meters. A plot of the data is given in Fig. 61.7. From this, it may be seen that there is a degree of spatial clustering in this data, for example, a group of higher-priced houses are visible to the north, and another group of lower-priced houses may be seen to the southern part of the area.

Here, an exponential variogram model is used – with a uniform prior on $[0, 30]$ km for h and a reciprocal prior for σ^2 . Calibration is via the MCMC approach, as outlined in the previous section. This is achieved using the *geoR* package for the R statistical programming language. As an example of the variogram calibration aspects of the analysis, the posterior distribution curve for the parameter h is shown in Fig. 61.8. Note that since *geoR* uses discrete approximations for distributions, this takes the form of a histogram. This shows that the posterior distribution for h peaks at around 4 km, although it has quite a long tail. In a classical inference approach, this kind of insight into inference about h is rarely provided, as in general it is just a two-number confidence interval that is provided.

In Fig. 61.9 the correlation derived from the variogram associated with each of the possible values for h is shown. For each h , the correlation curve is drawn with an intensity corresponding to the posterior marginal probability of that value of h . Reading off vertically from the x -axis suggests posterior probabilities of correlation associated with a given value of h . From this it can be seen that there is very little correlation between pairs of observations separated by more than about 20 km.

Fig. 61.7 Locations and asking prices for the Liverpool house price data

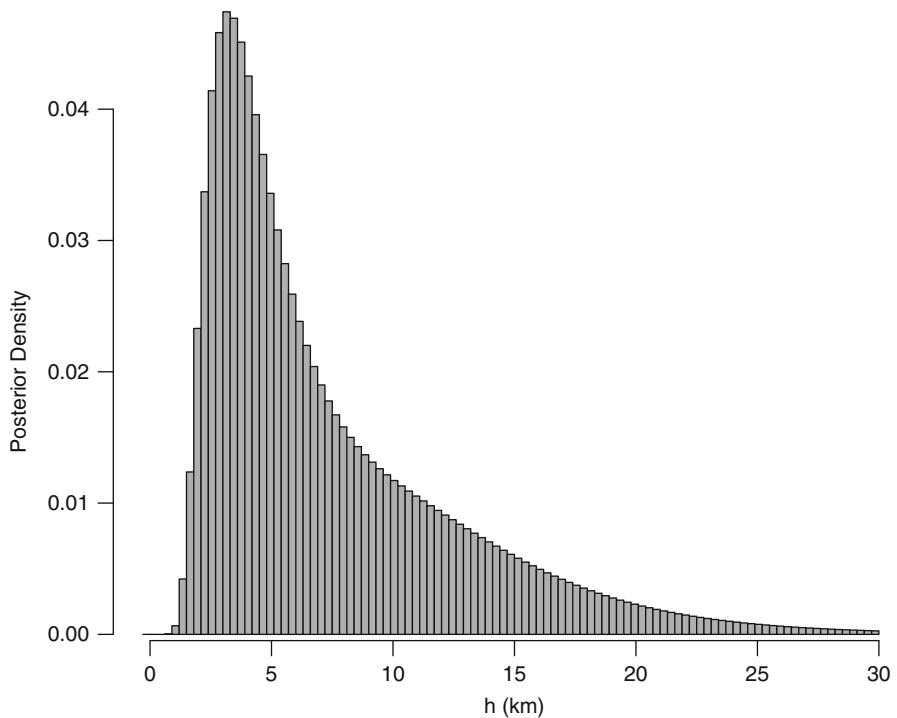
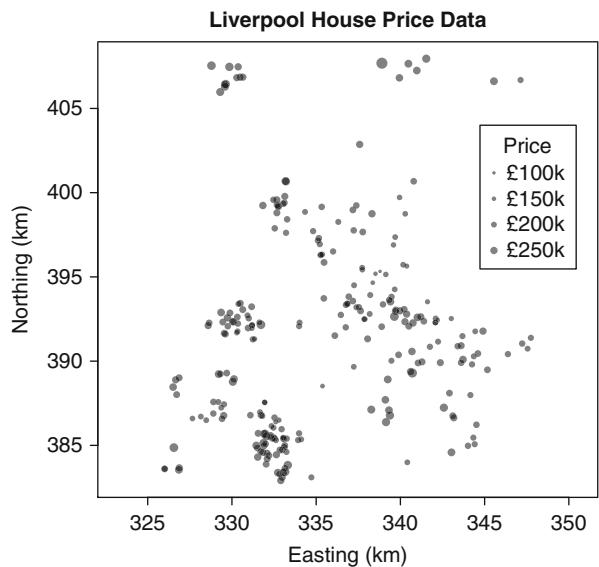


Fig. 61.8 Posterior distribution (via MCMC) of h in the house price variogram

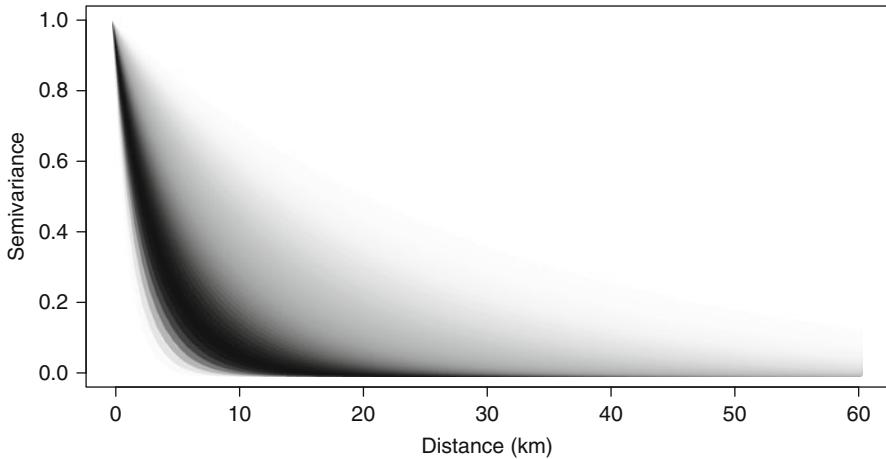


Fig. 61.9 Posterior estimates of the correlation curve

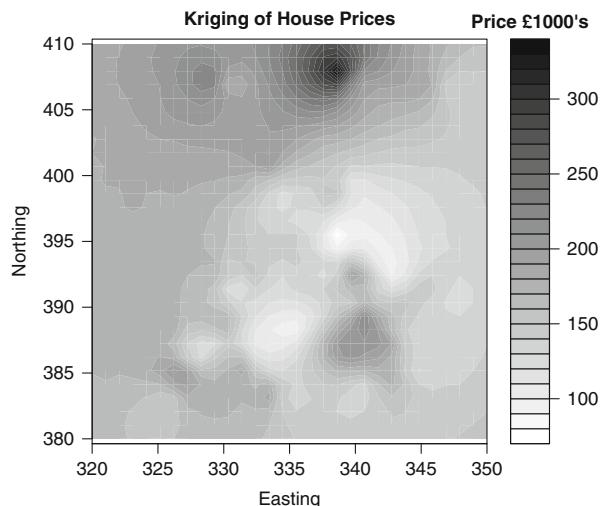
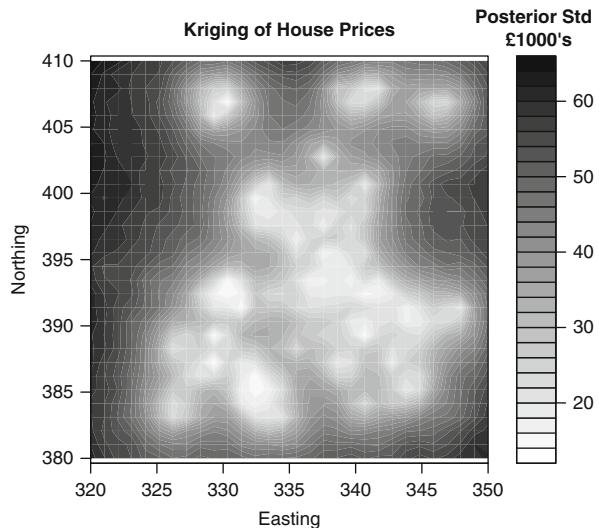


Fig. 61.10 Posterior predicted mean house price values based on Bayesian kriging

Moving on from inference about the variogram, as stated above, it is also possible to make inferences about the values of $E(z|\mathbf{x})$ where there are no observations. In this case, the technique is used to evaluate such values at regular points on a grid, so that “house price surfaces” can be drawn, outlining regional trends in house prices around the city of Liverpool. Again, the geoR package in R (Ribeiro and Diggle 2001) is used to achieve this. Results are shown in Fig. 61.10. As these are also the results of MCMC simulation, it is possible to visualize the accuracy of the estimates, by showing a corresponding surface of the standard deviations of the posterior distributions of the estimates of $E(z|\mathbf{x})$ at each location on the grid.

Fig. 61.11 Posterior predicted standard deviations of house price values based on Bayesian kriging



These are illustrated in Fig. 61.11. Note that in this figure, lower values of the posterior standard deviation (and hence greater confidence in the estimation) occur at locations close to the data points. To see this, compare Figs. 61.7 and 61.11.

61.6 Region-Based Measurement Data

In this section, data associated with regions (such as states or counties that may be represented by polygons) will be discussed. This section will be somewhat briefer than the previous two – a key reason for this is that the approach seen in Sect. 61.3 is quite similar to those that may be applied here. This is because, for point data, a discrete grid approximation was used, and this is essentially applying Bayesian approaches to region-based data, providing the “regions” form a regular lattice. A number of models have been developed to describe multivariate distributions where each variable is a quantity associated with a region. For example, if z_i is associated with region i , then the simultaneous autoregressive model (SAR) is

$$z_i = \mu_i + \sum_{j=1}^n b_{ij}(z_j - \mu_j) + \varepsilon_i \quad (61.28)$$

where $\varepsilon_i \sim N(0, \sigma_i^2)$ independently, and $\mu_i = E(z_i)$ and b_{ij} are constants, with $b_{ii} = 0$. Frequently, if w_{ij} is an indicator variable stating whether regions i and j are adjacent (with the convention that $w_{ii} = 0$), then the b -values are modeled as $b_{ij} = \rho_s w_{ij}$ and the values of μ_i and ρ_s are parameters to be estimated. Note that here the s is not intended as an indexing subscript; it simply denotes that this parameter is associated with the SAR model. In the simplest case for μ_i , we may assume it takes

a constant value μ and also that σ_i^2 takes the constant value σ^2 – so that all that is needed is to estimate the three values μ , σ^2 , and ρ_s .

A related model is the conditional autoregressive (CAR) model which specifies the distribution of z_i conditionally on all of the other z -observations, denoted by z_{-i} :

$$z_i|z_{-i} \sim N\left(\mu_i + \sum_{j=1}^n c_{ij}(z_j - \mu_j), \tau_i^2\right) \quad (61.29)$$

where, similar to before, $c_{ii} = 0$. Again, a common form of model for c_{ij} is $c_{ij} = \rho_c w_{ij}$. Assuming that the μ_i 's and τ_i^2 's are fixed for all regions, again we have a model with just three parameters, μ , τ^2 , and ρ_c . In each of these cases, Bayesian analysis can be carried out in a fairly intuitive way. For both the CAR and SAR models, it is possible to rearrange the model equations to the form of multivariate distributions of \mathbf{z} , the column vector of z_i observations. For the SAR model, we have

$$\mathbf{z} \sim N\left(\mu \mathbf{1}_n, \sigma^2 (\mathbf{I}_n - \rho_s \mathbf{W})^{-1} \left[(\mathbf{I}_n - \rho_s \mathbf{W})^{-1} \right]^T\right) \quad (61.30)$$

where $\mathbf{1}_n$ is a column vector of n 1's, \mathbf{I}_n is the $n \times n$ identity matrix, and \mathbf{W} is the matrix of w_{ij} values. Similarly the factorization theorem – see Besag (1974), for example – gives for the CAR model:

$$\mathbf{z} \sim N\left(\mu \mathbf{1}_n, \tau^2 (\mathbf{I}_n - \rho_c \mathbf{W})^{-1}\right) \quad (61.31)$$

Using similar prior distributions to those from Sect. 61.3 combine with either Eq. (61.29) or Eq. (61.28) to provide the likelihood functions enables a multivariate posterior distribution for σ^2 , ρ_s (or ρ_c), and μ to be derived and simulated. With this in mind, it is relatively straightforward to simulate posterior distributions for either of the sets of three distributions or for some priors to derive the distributions theoretically.

61.7 Conclusions

The above discussion demonstrates that it is possible to apply Bayesian inference to a number of spatial problems. Here, problems have been broadly classified in terms of the form of geographical information used as the basis of the analysis: point-based data, point-based measurement data, and region-based measurement data. In each case some form of Bayesian analysis is proposed. However, the list here is by no means exhaustive. For example, it may be possible to apply kriging and variogram-based methods (Oliver 2010) to regional data either by assigning a centroid point to each region or – perhaps more realistically – by expressing

regional values as the average or sum of points defined on a random field defined within that region. Also it is possible to extend these models – whereas in this chapter the expected value of the observed measurements has been modeled as a constant value, it is possible to adopt a regression approach and to express this quantity as a function of explanatory variables. Bayesian analysis provides a useful inferential tool for calibrating all of these models; they provide a relatively rich set of tools for drawing inference relating to model parameters and particularly when using MCMC tools, for drawing inferences relating to functions of these parameters, or predictive distributions for future observed variables. There are other forms of spatial data, and these may also be usefully analyzed using Bayesian techniques. For example, linear or network data have not been considered here, but such data forms may also be usefully analyzed using models which lend themselves to Bayesian analysis.

However, perhaps a greater challenge for spatial analysis is the ability to assess *which* model for a given data set is the most appropriate or whether a particular model is entirely inappropriate. The use of inferential tools as stated above is applied within a framework where the “true” model is a member of a family of models, for example, when analyzing regional data although the specific values of μ , ρ_c , and σ^2 may not be assumed, it is taken as given that the model can be specified in the general framework of a CAR model. However, how might one decide whether this is more appropriate than a SAR model? This is not a simple matter of testing “nested” models, when one set of models is a subset of another and inference can be based on posterior distributions of the parameters in the larger subset, but a matter of comparing structurally distinct models. Indeed, how does one choose between either of these models and a kriging-based model as suggested earlier. A number of ideas have been proposed for this in a Bayesian framework, but this area of research is relatively new; the *deviance information criterion* (DIC) of Spiegelhalter et al. (2002) is one attempt to address this – although some debate has been raised – see, for example, Ando (2007) who discusses a tendency for the DIC to favor models with larger numbers of parameters and proposes a modification to address this.

Another issue related to this is that of model appropriateness. Some recent criticism has been aimed at CAR and SAR based models – Wall (2004) notes that although the **W** matrix suggests that there is some spatial structure in the model, the actual variance-covariance matrices in Eqs. (61.30) and (61.31) can provide counterintuitive relationships between distances between regions and correlations, stating that

... although these covariances are clearly just functions of **B** or **C**, in general there is no obvious intuitive connection between them and the resulting spatial correlations.

(Here **B** and **C** are matrices of b_{ij} and c_{ij} coefficients). Although the Bayesian approach allows us to make relative assessments of the most appropriate parameter values within the modeling framework, it is also important to determine whether *any* combination of parameters would be meaningful. These observations are highly significant; a great deal of work (Bayesian and otherwise) has used models of this kind and has in general accepted that they have encapsulated spatial dependency in a reasonable way.

In summary then, it is argued that Bayesian analysis has a great deal to offer to spatial analysis and provides a richness of inferential tools, particularly via MCMC, that allow insights to be made that may not otherwise be possible or at least achieved as easily or intuitively. However, there are still a number of challenges, although arguably many of those – such as the issues of model appropriateness or model selection – are problems that face *all* kinds of statistical inference.

References

- Allen CD, Breshears DD (1998) Drought-induced shift of a forest–woodland ecotone: Rapid landscape response to climate variation. *Proc Natl Acad Sci* 95(25):14839–14842
- Ando T (2007) Bayesian predictive information criterion for the evaluation of hierarchical Bayesian and empirical Bayes models. *Biometrika* 94(2):443–458
- Anscombe FJ (1948) The transformation of Poisson, binomial and negative-binomial data. *Biometrika* 35(3–4):246–254
- Berry D (1997) Teaching elementary Bayesian statistics with real applications in science. *Am Stat* 51(3):241–246
- Besag J (1974) Spatial interaction and the statistical analysis of lattice systems (with discussion). *J R Stat Soc B* 36(2):192–236
- Besag J, Kooperberg C (1995) On conditional and intrinsic autoregression. *Biometrika* 82(4):733–746
- Cressie N (1993) Statistics for spatial data, Rev edn. Wiley, New York
- Diggle PJ, Tawn J, Moyeed R (1998) Model-based geostatistics. *Appl Stat* 47(3):299–350
- Fischer MM, Wang J (2011) Spatial data analysis. Models, methods and techniques. Springer, Berlin/Heidelberg/New York
- Holland M, Risser PG (1991) Ecotones: the role of landscape boundaries in the management and restoration of changing environments. Chapman and Hall, New York
- Hyndman RJ (1996) Computing and graphing highest density regions. *Am Stat* 50(2):120–126
- Mäkitalo M, Foi A (2011) A closed-form approximation of the exact unbiased inverse of the Anscombe variance-stabilizing transformation. *IEEE Trans Image Process* 20(9):2697–2698
- Mandalaz D (2008) Sampling techniques for forest inventories. Chapman and Hall/CRC, Boca Raton
- Matheron G (1970) The theory of regionalised variables and its applications. Les Cahiers du Centre de Morphologie Mathématique De Fontainebleau
- Matheron G (1973) The intrinsic random functions and their applications. *Adv Appl Probab* 5(3):439–468
- Oliver MA (2010) The variogram and kriging. In: Fischer MM, Getis A (eds) Handbook of applied spatial analysis. Springer, Berlin/Heidelberg/New York, pp 319–352
- Ribeiro P Jr, Diggle P (2001) geoR: a package for geostatistical analysis. *R-NEWS* 1(2):15–18
- Spiegelhalter DJ, Best NG, Carlin BP, van der Linde A (2002) Bayesian measures of model complexity and fit (with discussion). *J R Stat Soc Ser B Stat Methodol* 64(4):583–639
- Wall MM (2004) A close look at the spatial correlation structure implied by the CAR and SAR models. *J Stat Plan Inference* 121(2):311–324

Keith C. Clarke

Contents

62.1	Introduction	1218
62.2	Complexity and Models of Complexity	1219
62.3	Cellular Automata: Origins	1220
62.4	Cellular Automata: Key Contributions	1222
62.5	Cellular Automata: Applications	1223
62.6	Agent-Based Models: Origins	1225
62.7	Agent-Based Models: Key Contributions	1227
62.8	Agent-Based Models: Applications	1228
62.9	Conclusions	1230
	References	1231

Abstract

Two classes of models that have made major breakthroughs in regional science in the last two decades are cellular automata (CA) and agent-based models (ABM). These are both complex systems approaches and are built on creating microscale elemental agents and actions that, when permuted over time and in space, result in forms of aggregate behavior that are not achievable by other forms of modeling. For each type of model, the origins are explored, as are the key contributions and applications of the models and the software used. While CA and ABM share a heritage in complexity science and many properties, nevertheless each has its own most suitable application domains. Some practical examples of each model type are listed and key further information sources referenced. In spite of issues of data input, calibration, and validation, both

K.C. Clarke

Department of Geography, University of California, Santa Barbara, Santa Barbara, CA, USA
e-mail: kclarke@geog.ucsb.edu

modeling methods have significantly advanced the role of modeling and simulation in geography and regional science and gone a long way toward making models more accountable and more meaningful at the base level.

62.1 Introduction

Models are simplifications of real-world systems that are amenable to tests and simulations of the reactions of the real systems to changes in their state and function. For extant and complicated regional systems, such as the United States Interstate Highway system, experiments on society would be unacceptable (closing highways to measure traveler delays, for example) – yet the computer allows such experiments *in silico*. Models, of course, are only of value if their structures are based on knowledge or data about an actual system and if they give results which are reasonable and credible. Foremost among the challenges of modeling is the fine-tuning of models so that they achieve the best results (calibration), of meaningfully converting a system’s components into structural and behavioral equivalents within the model (design), of the model’s effective use of computing power (tractability), of the ability to match actual or expected results (performance), and of their ability to create accurate predictions (validity). Regional science has employed a large number of modeling approaches over time, yet in the last three decades, two paradigms of modeling have emerged that have made achievements against these challenges and that have led to breakthroughs in model performance and accuracy for regional systems. These two approaches are cellular automata (CA) models and agent-based models (ABM).

In this chapter, we examine these two modeling approaches. Both have been termed “individual-based modeling approaches” in ecology, and this reflects the fact that both types of models are bottom-up – that is, they model the primitive or elemental level of behavior associated with a system. Aggregate patterns are achieved by summing the results of many individual actions, which has led to the related terms “disaggregated models” and “micro-simulation models.” Both these approaches are similar in that they are simple, easy to program and implement, and use an iterative approach. Both require initial conditions to be set and have challenges around calibration procedures. Cellular models are preferred when geographic space can be represented in the form of a geographic grid, such as the cells in a raster Geographic Information System. They are also favored when model states and the probabilities of transitions among those states are known and stable. They are most suitable for dissipative processes, such as land use change and urban growth. On the other hand, agent-based models are superior when the basis of a model is a behavioral unit, such as a person, household, business, landholder, or farmer (the “agent”), and when the modeled process consists of interactions over time among one or more types of agents that produce a spatial form, such as land use, crop choice, or habitat type. It has been said that the two modeling forms differ only in the fact that in CA the agents remain in place and interact only with their neighbors. This statement, however, ignores

both major and subtle differences between the two modeling approaches, such as their means of calibration and validation. We return to this contrast in the concluding section.

62.2 Complexity and Models of Complexity

Complex systems theory was originally developed in physics and has origins in Lorenz's work on weather forecasting, which in turn reflect chaos theory and work on the three-body problem by Poincaré in 1890. Initially, Lorenz observed that a system's behavior in the long term reflects the initial conditions of the system, such as the locus of an attraction point being a function of where a point subject to the attraction started its path. The values of variables that separate different system behaviors are called thresholds, and crossing them leads to nondeterministic and nonlinear behavior. Complexity is that behavior phase which is neither static nor deterministic. An early demonstration of complexity was in sand piles. When sand is poured from a nozzle, it forms a pile, which grows in a simple linear fashion. However, at some point in its growth, the sides of the pile are subject to failure. Even though the exact failure gradient is known, it is impossible to tell when a failure will take place and how much of the sand pile it will take down. Such behavior has been called self-organized criticality.

As chaos and complexity theory became more known, largely due to the Santa Fe Institute and the work of scholars like Murray Gell-Mann and John Holland, applications in many different fields became commonplace. Complexity has a natural link to the science of fractals and self-similarity, as noted by Batty (2000). Many of the fields that adopted the complex systems approach were related to physical geography, such as meteorology, fire modeling, and ecological succession. However, Batty and Longley's (1994) demonstration of the fractal nature of cities led to some degree of acceptance within urban and human geography. Many systems in human geography exhibit complexity, including land use change, residential segregation, urban growth, road network growth, and intercity interactions.

Important concepts in complexity theory are that dynamical systems – those subject to feedbacks – exist in three aggregate states or phases: chaos, stability, and complexity. In chaos, no discernable rules, structures, or even heuristics apply, such as in the business cycle or the stock market. In stability, behavior is linear or can be modeled by polynomials, that is, the change is differentiable and solvable with differential equations, equilibrium theory, and optimization. Complexity, however, is marked by periods (time) or subregions (space) of both stability and chaos. A system can move from one aggregate behavior state to another (a phase change), but each behavior type is robust (resilient) against perturbation to some degree (Waldrop 1993). Tipping a system beyond a threshold provokes a phase change, and the system then trends away from the original state. An example often used is a lake, which is subject to inputs of phosphates. The ecosystem of the lake is able to counter the impact of the phosphates up to a certain concentration. Beyond that, even by a fraction, the lake cannot return to its initial state, and eutrophication takes

place, leading to a new ecosystem based on the higher phosphate levels and different plant and animal species. An unknown, possibly large, proportion of human and natural systems exhibit such complexity.

The attraction of both cellular automata and agent based models is they represent some of the simplest frameworks possible for demonstrating complex systems behavior. Largely for this reason, the models were quickly adopted and used to test many new types of urban and economic systems models. John Holland has suggested a defining condition for identifying complex systems and complexity, which he has termed emergence (Holland 1998). Emergence has been criticized as too subjective a criterion by which to indentify complexity but is said to exist in a system when new and unpredicted patterns or global-level structures arise as a direct result of local-level procedures. The structure or pattern that emerges cannot be understood or predicted from the programmed or assumed behavior of the individual units alone. An example of emergence in CA is the glider (see Sect. 62.3). An example in the SLEUTH CA urban model (Clarke et al. 2007) is the aggregation of new settlements at the junctions of roads, a behavior nowhere inherent in the model's programmed behavior.

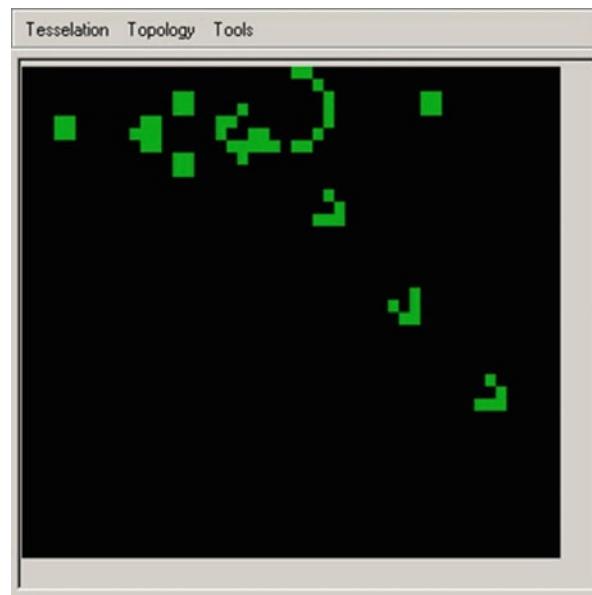
62.3 Cellular Automata: Origins

A cellular automaton (plural cellular automata) is a discrete model originally theoretical, but now implemented in disciplines from physics to biology, geography to ecology, and computer science to regional science. CA have been defined as “discrete spatio-temporal dynamic systems based on local rules” (Miller 2009). As noted above, they are the simplest modeling framework in which complexity can be demonstrated. Using CA, extremely complex behavior and emergence can be demonstrated with terse conditions and minimal rules. They are inherently attractive as spatial models because they map closely onto the raster grid in a geographic information system, because they use only local interactions among cells, and because of their simplicity. Nevertheless, they are capable of modeling and simulating extraordinarily complex behavior (Batty 2000) and of demonstrating emergence.

A CA has four elements: (i) a grid of cells, each of which can assume a finite number of states; (ii) a neighborhood, over which a change operator applies, usually the Moore (8-cell) neighborhood surrounding a cell in the grid; (iii) a set of initial conditions, that is, an instance of the states for each and every cell in the system; and (iv) one or more rules, which when applied change the state of a cell based on properties or states of the neighborhood cells. The model advances by applying the rules to every cell one at a time, then swapping the changed grid with the initial grid, and by repeating this procedure.

CA were invented by Stanislaw Ulam, while he was employed at the Los Alamos National Laboratory in the 1940s. At the same time, John von Neumann was working on the problem of self-replicating systems. Von Neumann proposed the kinematic model, a robot that could rebuild itself from spare parts. Ulam recommended that von

Fig. 62.1 A 2D cellular automaton Game of Life. Configuration shows a glider gun, a cell form that remains static and sends out streams of gliders. Gliders and glider guns are emergent behavior from the simple rules of the Game of Life



Neumann developed his idea around a mathematical abstraction, such as the one he was using to study crystal growth on a lattice network. Like Ulam's lattice network, von Neumann's cellular automata used a two-dimensional grid, with his self-replicator implemented algorithmically, working within a CA with a 4-cell neighborhood and with 29 states per cell. This CA is now termed a von Neumann universal constructor. At about the same time, Norbert Wiener and Arturo Rosenblueth developed a CA model and mathematical description of impulse conduction in cardiac systems, implying broad applicability of the theory. By the 1960s, CA were being studied as a simplification of dynamical systems – models developed to simulate natural systems with feedbacks, such as air flow, turbulence, and weather, and human systems such as cities and economies. In 1969, Gustav Hedlund compiled many CA results into a seminal paper on the mathematics of CA (Hedlund 1969).

Nevertheless, CA remained largely a mathematical curiosity until John Conway's creation of a CA game, the Game of Life. Martin Gardner drew popular attention to the game in a 1970 issue of his games column in *Scientific American* (Gardner 1970). Life was a two-state, two-dimensional CA with only four rules: (i) Any live cell with fewer than two live neighbors dies (death), (ii) Any live cell with two or three live neighbors remains alive (survival), (iii) Any live cell with more than three live neighbors dies (overcrowding), (iv) Any dead cell with exactly three live neighbors becomes alive (birth). Despite the game's simplicity, it can create astonishing variety in its long-term patterns. An “emergent” phenomenon is the “glider,” a cell arrangement that perpetuates itself by continuous movement across the grid (Fig. 62.1). It is possible to arrange the automata so that gliders interact to perform computations, and it has been proven that the Game of Life can emulate a universal Turing machine, thus completing von Neumann's line of research.

62.4 Cellular Automata: Key Contributions

In the 1980s, Stephen Wolfram published a series of papers systematically investigating an unknown class of one-dimensional cellular automata, which he called *elementary cellular automata* (Wolfram 1986). The demonstration that what is now termed “complex systems behavior” can be simulated from the simplest of CA led to a host of explorations within the social and physical sciences into the range of what CA could simulate. Wolfram continued this work, and in 2002 published *A New Kind of Science* (Wolfram 2002). In the book, Wolfram argues that discoveries about cellular automata are not isolated facts but have significance for all disciplines of science. Using a one-dimensional CA, Wolfram demonstrated that virtually any mathematical function can be simulated, and he explored applications across disciplines. Wolfram proposed a four-class set of possible CA. In Class 1, nearly all patterns quickly evolve into a stable homogenous set and randomness disappears. In Class 2, nearly all patterns quickly evolve into an oscillating structure, with some randomness remaining. In Class 3, nearly all patterns evolve into pseudo-random or chaotic structures. Any regular structures are quickly eliminated by randomness, which dissipates through the entire system. In Class 4, nearly all initial patterns evolve into structures that interact in complex and interesting ways. Wolfram has conjectured that many class 4 cellular automata are capable of universal computation. This has been proven for Conway’s Game of Life and for Wolfram’s Rule 110. Rule 110 is a unique achievement, defined as a one-dimensional CA that for the input neighboring configuration set {111, 110, 101, 100, 011, 010, 001, 000} yields the equivalent outputs {0,1,1,0,1,1,1,0}. Of the 88 possible unique elementary cellular automata, Rule 110 is the only one for which Turing completeness has been proven, making it arguably the simplest known Turing complete system. Rule 110 exhibits Class 4 behavior, which is neither completely stable nor completely chaotic. Localized structures appear and interact in various complicated-looking ways, demonstrating the properties of emergence and phase change. There have been several attempts to place CA into other formally rigorous classes, inspired by Wolfram’s classification. For instance, Culik and Yu proposed three well-defined classes (and a fourth one for the automata not matching any of these) called Culik-Yu classes.

From the perspective of geocomputation, Batty (2000) surveyed the variants of CA possible for simulating urban and similar systems. He pointed out that strict CA models are on one end of a computational spectrum and that at the other end are simple Cell Space models, really no different than raster grids with a finite set of states that transition over time. He distinguished between cell space models, which are not at all CA models in the strict sense, and the concept of relaxing the CA assumptions. Key among the relaxations is the incorporation of action-at-a-distance, which is excluded by strict CA’s use of the von Neumann or Moore neighborhoods only. CA development in modeling of urban areas and other geographical realms are covered in a literature review, and some useful information sources are listed (Batty 2000, p. 119).

Beyond mathematics, CA applications have been less concerned with definitional rigor and more with making CA adjust to geographical variation. Approaches have included automatic learning methods to empirically derive rules from observed patterns, self-modification or rule changes triggered by aggregate system behavior, and the addition of “ghost” states, that fall between strict classes (e.g., “urban,” “non-urban,” and “under development” as land uses). Sante et al. (2010) note eight ways that formal CA models have been modified for use in urban growth modeling: using irregular spaces, nonuniform cell spaces, extended neighborhoods, nonstationary neighborhoods, complex transition rules, nonstationary transition rules, by adding growth constraints, and using irregular time steps. Theoretical work has also examined the synchronous versus asynchronous application of the rules.

Applications of CA models in regional science have been commonplace. Most frequently, the models work on land use maps, often simplified to urban and nonurban states. Rules are derived and models calibrated using past data states, that is, by hindcasting. Land use maps derived from remotely sensed data at different time periods have commonly been used as data inputs, and other data are often zoning restrictions, transportation networks, and topography. Geographic Information Systems are used to compile and georegister the map layers, and to receive the modeling results. CA models have been applied at many scales, following the research on fractal urban forms pioneered by Michael Batty (Batty 2005), but most CA models use data at resolutions between 30 and 100 m. An early model by White and Engelen (1993) added action-at-a-distance by changing the Moore neighborhood assumption. Clarke et al. (1997) created the SLEUTH model, a CA that incorporated weighting of probabilities and self-modification, feedback from the aggregate to the local. Wu and Webster’s modeling of the rapid growth in Southern China was another significant contribution (Wu and Webster 1998). Sante et al. (2010) tabulated 33 urban CA models and compared their characteristics, and provided a useful summary of the theoretical and applied CA modeling surrounding geography, urban planning, and regional science. Silva has considered complexity theory in planning more generally, using CA as the specific example (Silva 2010).

Sante et al. (2010) also offered a classification of CA transition rules. Type I rules are those of classical CA, that is, transitions can only occur based on the states of neighboring cells. Type II rules are based on potentials or probabilities altered by the land or environmental status of a cell. Type III rules are pattern development rules, which adjust the states based on shape or the existence of a network, such as roads. Type IV rules use computational intelligence methods to determine the rules from prior system behavior. Typical are Case-Based Reasoning, neural networks, data mining and kernel-based methods. Type V rules use fuzzy logic and uncertainty reasoning, while Type VI rules include those not compatible with types I–V.

62.5 Cellular Automata: Applications

Examples of cellular models in popular use include DINAMICA (See: www.csr.ufmg.br/dinamica), SLEUTH (See: www.ncgia.ucsb.edu/projects/gig), and

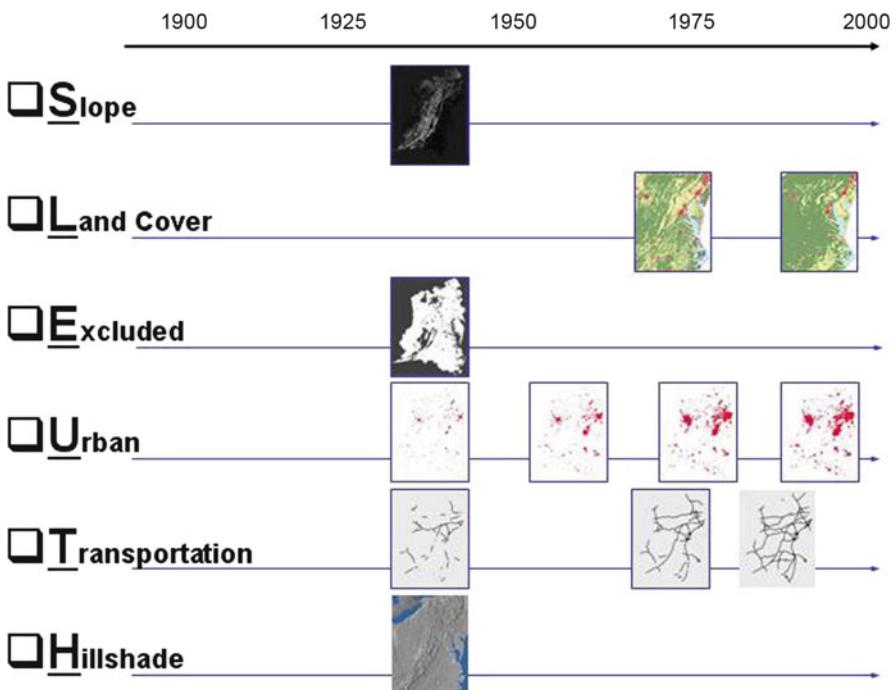


Fig. 62.2 Input data for the SLEUTH model. Minimum layers are topographic slope, two land use maps, one exclusion map, four urban extent maps, two transportation maps, and a hill-shaded background image. Data shown are for the Environmental Protection Agency's Mid Atlantic Regional Assessment study area

Simland (Wu 1998). Influential critical reviews of CA research include those by Batty (2005), Torrens and O'Sullivan (2001), and Benenson (2007). An important early theoretical framework was that of Takeyama and Couclelis (1997), and additional attempts at synthesis have been made by Benenson and Torrens (2004) and by Torrens and Benenson (2005). Nevertheless, interest in and use of the CA suite of models continues unabated, with applications of several of the models at many scales, across regions, for whole nations, and on all continents other than Antarctica.

A representative CA model that has been long-lived in relation to others is the SLEUTH model. SLEUTH is an acronym for the data input layers required by the model (Fig. 62.2). The model was developed by the author and a host of collaborators with funding support from the United States Geological Survey, the National Science Foundation, and the Environmental Protection Agency. There are three retrospectives on SLEUTH's now 15 years of use (Clarke et al. 2007; Clarke 2008a, b). SLEUTH actually consists of two CA models tightly coupled together and coded within the same open source C-language program: the Clarke Urban Growth model and the Deltatron Land Use Change model.

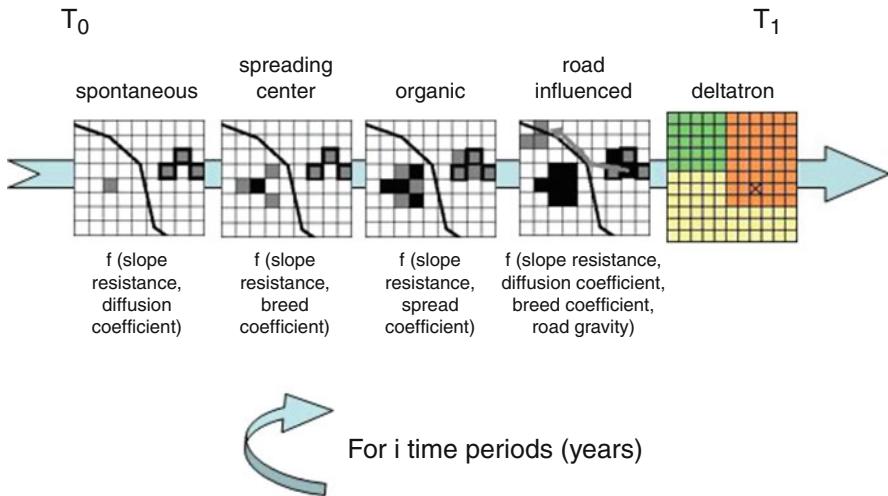


Fig. 62.3 At each cycle in the CA model, five sets of behavior rules are enforced. These are controlled by the factors and parameters shown and are applied in sequence for each one “year” iteration of the model

The former is a classic CA, using a Moore neighborhood and simple sequential rules (Fig. 62.3), but using weighting for probabilities, Monte Carlo simulation, and self-modification – in which aggregates, such as the overall growth rate, feedback into the parameters controlling the rule sets. The latter differs in that it takes its input of quantity of transformation from the Urban Growth model and applies CA in change space rather than geographic space. In doing so, it relaxes the single time-step rule and allows persistence and aging of cells for longer than one time step. SLEUTH has over a hundred applications at many scales and for different cities worldwide. A typical forecasting result is shown in Fig. 62.4.

62.6 Agent-Based Models: Origins

Agent-based models (ABM) are a class of computational models for simulating the actions, behavior, and interactions of autonomous individual or collective entities, with the goal of exploring the impact of one agent or a behavior type on the system as a whole. Miller (2009) notes that the agents are independent units that attempt to fulfill a set of goals. The agents can be countries, landowners, residents, renters, farmers, shoppers, vehicles, or even people out for a walk. Unlike with CA, the purpose of ABM is often the exploration of variants in system behavior due to agent characteristics (such as the proportion of agents of different types) or rules, rather than resulting aggregate structures or maps. Multiagent models include more than one agent; for example, a habitat model may include plants, animals that eat the plants, and predators that eat the animals. ABMs combine game theory, complex

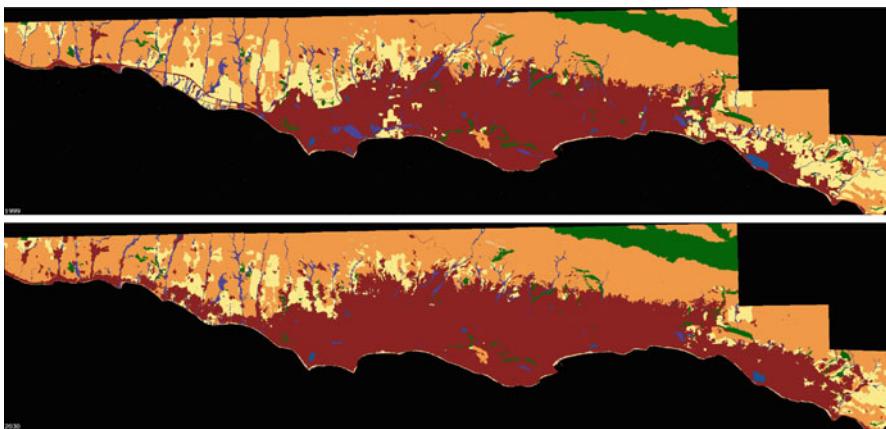


Fig. 62.4 Land use in the Santa Barbara, California, region. *Top:* In 1998, the base year for modeling with SLEUTH. *Bottom:* In 2030, as forecast by SLEUTH for the Santa Barbara Regional Impacts of Growth Study (2003). *Black:* unclassed; *Red:* urban; *yellow:* agriculture; *orange:* rangeland; *green:* forest; *blue:* water; *purple:* wetland; *Tan:* barren land

systems theory, evolutionary programming, and stochastic modeling. In ecology and biology, ABMs are termed “individual-based models.”

ABMs simulate the actions and interactions of multiple agents, in an attempt to emulate the overall system behavior and to predict the patterns of complex phenomena. Agents behave independently, but react to the environment, the aggregate properties of the system, and other agents. So, for example, a farmer agent in the Brazilian Amazon may clear land as he becomes more profitable, in response to a change in crop price, or because his neighbor is clearing his land. Agents are usually assumed to behave with bounded rationality, acting in their own interests such as reproducing, increasing profit, or increasing status, usually by simple heuristic rules. For example, the previously mentioned farmer may decide to have a child or build a house when profitability reaches a certain level. Agents can also “learn,” that is, avoid previously failed decisions while favoring successful ones. They can also adapt, that is, change behavior based on properties of the system.

An ABM consists of (i) agents specified at specific model scales (granularity) and types; (ii) decision-making heuristics, often informed by censuses and surveys in the real world; (iii) learning or adaptive rules; (iv) a procedure for agent engagement, for example, sample, move, interact; and (v) an environment that can both influence and be impacted by the agents. Creating a model involves examining or surveying a system to extract the agents’ behavior and influential factors, quantifying these elements, then coding the model in an environment that allows control, examination of maps and time sequences, and metrics of system behavior and performance. Many ABMs are programmed in coding languages with Java being the most common, while others use one or more of the software tools,

both open source and proprietary, in which the system and rules have to be specified. While there are many examples of software for ABM, relatively few of them are compatible with GIS or produce maps or images. Also of use is the ability to do Monte Carlo simulations and to let the models iterate to a steady state.

ABMs share their origins with CA in the work of von Neumann, Ulam, and Conway. A pioneering agent-based model in urban systems was Thomas Schelling's urban residential segregation model (Schelling 1971). Though not computational, the work embodied the basic concept of agent-based models as autonomous agents interacting within a fixed environment and with an observed aggregate outcome. In the 1980s, interest in game theory led to Robert Axelrod's experiments with the game "Prisoner's Dilemma," showing that strategies evolved and coevolved over time among players. Craig Reynolds' research on models of flocking behavior yielded the first biological agent-based models with embedded social characteristics. Modeling biological agents using ABM became known as artificial life. This led to artificial societies, artificial cities, computational economics, etc. Important software tools for ABM were StarLogo, SWARM, and NetLogo in the 1990s, and since then Ascape, Repast, Anylogic, and MASON (Railsback et al. 2006). Examples of early models include *Construct* by Kathleen Carley and *Sugarscape* by Joshua Epstein and Robert Axtell. These explored the coevolution of social networks and culture and the role of social phenomena such as segregation, migration, pollution, sexual reproduction, combat, and the transmission of disease. An early book on ABM in social simulation was Nigel Gilbert's *Simulation for the Social Scientist* (1999). A key research journal has been the *Journal of Artificial Societies and Social Simulation*.

62.7 Agent-Based Models: Key Contributions

A survey of the recent ABM literature is that of Niazi and Hussain (2011). A key survey in Geography was that of Parker et al. (2003), resulting from a workshop (Parker et al. 2002). Also influential was a series of papers published by the Santa Fe Institute (Gamblett 2002). Agent-based modeling has been extraordinarily interdisciplinary. ABM has been applied to model organizational behavior, logistics, consumer behavior, traffic congestion, building and stadium evacuation, epidemics, biological warfare, and population demography. In these cases, a system encodes the behavior of individual agents and their interconnections. In some geographical applications, the models have been informed by field work, interviews, or from censuses that are used to derive behavioral characteristics and choices using qualitative methods. Agent-based modeling tools are then used to test how changes in individual and collective behavior impact the system's aggregate behavior. In some cases, agents are allowed to learn from past choices, avoiding decisions with negative outcomes, for example.

The following ABM development environments include the ability to ingest, output, and use spatial data: Anylogic, Cormas, Cougaar (via OpenMap),

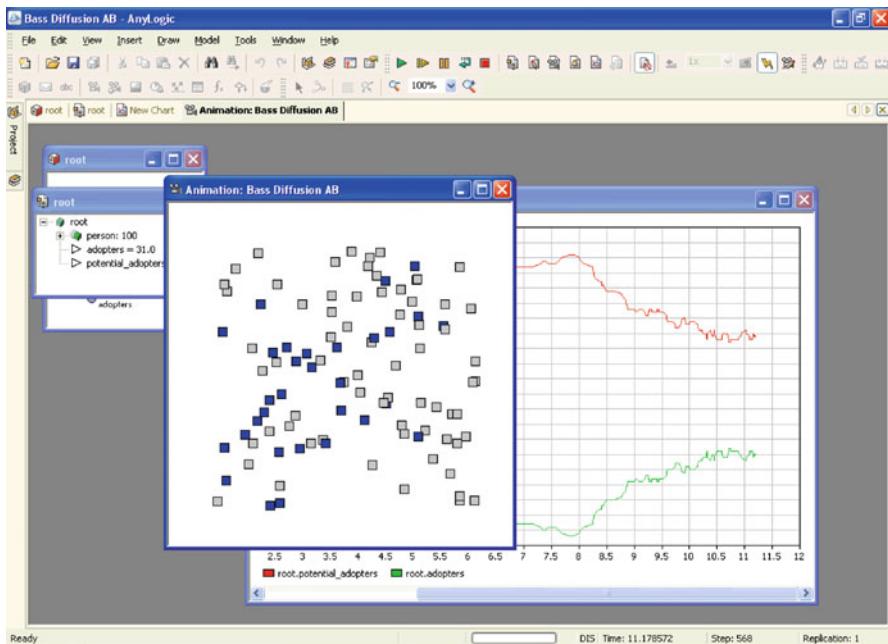


Fig. 62.5 User Interface for the Anylogic 6 Agent-Based modeling software (Source: http://www.coensys.com/agent_based_models.htm)

Framsticks, Janus (using JaSIM), MASON, Repast, SeSAM, NetLogo, and VisualBots. Some of these, and other nonspatial packages, contain model libraries that include CA examples such as Game of Life, HeatBugs, demographic models, epidemiological models, and flocking. Many include the means to display charts, graphs, and maps and menus to input control variables and rules (Fig. 62.5).

62.8 Agent-Based Models: Applications

Recent topics in regional science that have been modeled with ABMs include crowd behavior during riots and outdoor events (Torrens 2012), innovation in businesses (Spencer 2012), commuting behavior (McDonnell and Zellner 2011), ecology and habitats, disease, and land use change. The most recent research on agent-based models has demonstrated the need for combining agent-based and complex network-based models. This has included a desire for models with reusable components, tools for proof of concept and design, descriptive agent-based modeling for developing descriptions of agent-based models by means of templates and complex network-based models, and a need for better validation. The latter point has been repeatedly used in critiques of ABM: their very nature makes

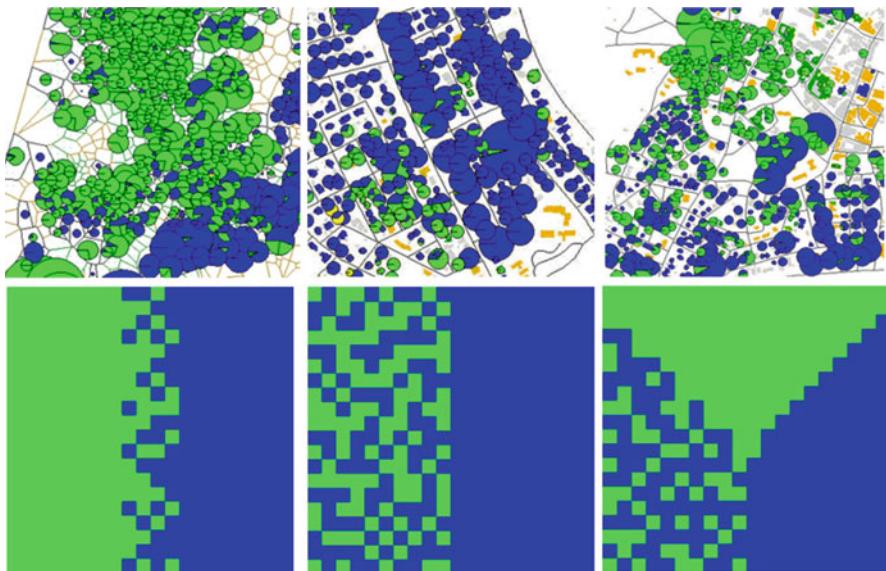


Fig. 62.6 Israeli and Arab residential patterns of the Israeli towns of Yaffo and Ramle from the Israeli census, and their Schelling model simulations (Source: Hatna and Benenson (2012))

calibration and validation using data descriptive of real systems rather difficult, if not impossible. For example, a programmed behavior type will indeed emerge when enough agents are given enough time, so experiments with agents could be considered circular reasoning. Truly emergent behavior or new knowledge should be unanticipated during model construction.

An example of a common ABM application is the Schelling model (Schelling 1971). This simple model of segregation, originally proposed as a game simulation, has been used as the basis of many agent-based models and theoretical discussions about ABMs. The model illustrates how individual tendencies regarding neighbors can lead to segregation in cities. The model has been extensively used to study residential segregation of ethnic groups where agents represent householders who relocate in the city. A concise statement of the model is that by Benenson et al. (2009), in which are enumerated the six behavior rules, assuming the model to be an ensemble of agents of two types, B and W, dispersed over a grid. The rules are: (i) at every iteration, the agents can move to a vacant cell; (ii) the decision to move to a cell is based on the fraction in the neighboring cells of the opposite type of agent; (iii) this fraction should be below a tolerance proportion; (iv) if this fraction is exceeded, then an agent moves; (v) the agent searches within a finite distance for cells that are below the tolerance threshold, and if none exists, does not move; (vi) vacated cells become available for other agents. In most applications, these rules produce residential segregation, depending on the two constants that must be determined. Benenson and his colleagues have repeatedly experimented with

real-world data and the Schelling model. [Figure 62.6](#) shows Jewish-Arab residential patterns of the Israeli towns of Yaffo and Ramle from the Israeli census, and their ABM equivalents ([Hatna and Benenson 2012](#)).

62.9 Conclusions

Cellular automata and agent-based models have both represented a new approach in modeling, that of complex adaptive systems. In this approach, models are micro-simulations, run at the atomic level, and aggregate behavior emerges as a consequence of large numbers of agent interactions. The complex systems approach has favored CA and ABM over Forrester-type systems dynamics models and steady-state and equilibrium models. CA and ABM share in common their individual basis. In CA, the modeled entities are cells that remain static while spatial and other processes move across or through them. In ABM, the agents can move in space, interact with each other directly, and interact with other agent types. In both cases, a large number of independent autonomous lowest level actors create the overall landscape. The models can include extra data, such as environmental control layers, and parameters that influence the agents, such as prices or demands.

CA models have been criticized as oversimplifications of reality, and those that have relaxed the rules have been criticized as not pure CA. CA are extremely sensitive to their initial conditions, and are very consumptive of CPU time. Since most use square grids, they are subject to error due to incorrect choice of map projection, and directional bias. In the long run, they are subject to equifinality arguments, since in most cases all developable land becomes developed, regardless of the exact sequence of development. They have also been criticized as difficult to implement and data hungry.

Bithell et al. ([2008](#), p. 625) have noted that ABM have the potential to create integrated models that cross disciplines, so that similar computational methods can be used to control the spatial search process, to deal with irregular boundaries, and display the changing of systems where the “preservation of heterogeneity across space and time is important.” They note that a principal challenge of ABM is to find sets of rules that best represent the beliefs and desires of human as agents, so that they reflect the cultural context, yet still allow system exploration. Clifford ([2008](#), p. 675) noted that ABMs are most appropriate where decisions or actions are distributed around specific locations, and where structure is seen as emergent from the interaction among individuals. For this new and exploratory modeling framework, he calls for “a rediscovery and reappraisal of the richness and depth of insight in the model-building enterprise more generally.” Some have attempted to link ABM with other bodies of theory, for example, Neutens et al. ([2007](#)) have linked ABM and time-space geography. Andersson et al. ([2006](#)) have attempted to link networks, agents, and cells to model urban growth. Lastly, O’Sullivan and Hakley ([2000](#)) have suggested that using ABM encourages a modeling bias toward an individualist view of the social world, thereby missing many forces that shape

real economic and human systems top down, such as planning and government. Read (2010, p. 329) noted that agent-based models “sometimes provide only a veneer of, rather than substantive engagement with, social behavior.”

Both CA and ABM have enjoyed popularity in the regional modeling arena in the last 20 years due to their simplicity, ease of use and accuracy. When machine learning or optimization is involved, the models can produce simulations that are of excellent accuracy. However, the models are often only as good as the data with which they are trained or tested, and are highly sensitive to the context of these data. Relatively few CA or ABM models are highly ported across different applications. Even fewer have been rigorously tested for accuracy, repetition, and parameter sensitivity and validated using independent data. A major criticism of both model types is that while the simulations are accurate and engaging, they lack any causative description or policy-related link to actual system behavior. Thus, while they can create useful future scenarios or forecasts, the means by which the actual system can be steered toward that outcome is not forthcoming. CA models are best used for spatially distributed process simulation, such as spread and dispersal, and when the geometry, scale, and basic behavior of a system are known. ABM is suited to simulations with no prior precedent, no past data, or when system knowledge is absent. These applications are usually more exploratory than when CA are used. Nevertheless, both modeling methods have significantly advanced the role of modeling and simulation in regional science and gone a long way toward making models more accountable and more meaningful at the base level. Their joint impact on research and understanding of human systems has been profound.

References

- Andersson C, Frenken K, Hellervik A (2006) A complex network approach to urban growth. *Environ Plan A* 38(10):1941–1964
- Batty M (2000) Geocomputation using cellular automata. In: Openshaw S, Abrahart RJ (eds) *GeoComputation*. Taylor and Francis, London, pp 95–126
- Batty M (2005) Cities and complexity: understanding cities with cellular automata, agent-based models, and fractals. MIT Press, Cambridge, MA
- Batty M, Longley P (1994) *Fractal cities: a geometry of form and function*. Academic Press, San Diego, CA and London
- Benenson I (2007) Warning! The scale of land-use CA is changing! *Comput Environ Urban Syst* 31(2):107–113
- Benenson I, Torrens PM (2004) Geosimulation: automata-based modeling of urban phenomena. Wiley, New York
- Benenson I, Erez H, Ehud O (2009) From Schelling to spatially explicit modeling of urban ethnic and economic residential dynamics. *Sociol Method Res* 37(4):463–497
- Bitell M, Brasington J, Richards K (2008) Discrete-element, individual-based and agent-based models: tools for interdisciplinary enquiry in Geography? *Geoforum* 39(2):625–642
- Clarke KC (2008a) Mapping and modelling land use change: an application of the SLEUTH model. In: Pettit C, Cartwright W, Bishop I, Lowell K, Pullar D, Duncan D (eds) *Landscape analysis and visualisation: spatial models for natural resource management and planning*. Springer, Berlin, pp 353–366

- Clarke KC (2008b) A decade of cellular urban modeling with SLEUTH: unresolved issues and problems, Ch. 3. In: Brail RK (ed) Planning support systems for cities and regions. Lincoln Institute of Land Policy, Cambridge, MA, pp 47–60
- Clarke KC, Hoppen S, Gaydos L (1997) A self-modifying cellular automaton model of historical urbanization in the San Francisco Bay area. *Environ Plan B Plan Des* 24(2):247–261
- Clarke KC, Gazulis N, Dietzel CK, Goldstein NC (2007) A decade of SLEUTHing: Lessons learned from applications of a cellular automaton land use change model, Chapter 16. In: Fisher P (ed) Classics from IJGIS. Twenty years of the International Journal of Geographical Information Systems and Science. Taylor and Francis/CRC Press, Boca Raton, pp 413–425
- Clifford NJ (2008) Models in geography revisited. *Geoforum* 39(2):675–686
- Gardner M (1970) Mathematical games: the fantastic combinations of John Conway's new solitaire game "life". *Sci Am* 223(9):120–123
- Gimblett HR (2002) Integrating geographic information systems and agent-based modeling techniques for simulating social and ecological processes. *Institute Studies in the Sciences of Complexity*, Oxford University Press, Santa Fe
- Hatna E, Benenson I (2012) The Schelling model of ethnic residential dynamics: beyond the integrated – segregated dichotomy of patterns. *J Artif Soc Soc Simul* 15(1):6
- Hedlund GA (1969) Endomorphisms and automorphisms of the shift dynamical system. *Math Syst Theory* 3(4):320–375
- Holland JK (1998) Emergence: from chaos to order. Addison-Wesley, Redwood City, CA
- McDonnell S, Zellner M (2011) Moira exploring the effectiveness of bus rapid transit a prototype agent-based model of commuting behavior. *Transp Policy* 18(6):825–835
- Miller HJ (2009) Geocomputation. In: Fotheringham AS, Rogerson PA (eds) *The SAGE handbook of spatial analysis*. Sage, London, pp 397–418
- Neutens T, Witlox F, Van de Weghe N, De Maeyer PH (2007) Space-time opportunities for multiple agents: a constraint-based approach. *Int J Geogr Inf Sci* 21(10):1061–1076
- Niazi M, Hussain A (2011) Agent-based computing from multi-agent systems to agent-based models: a visual survey. *Springer Scientometr* 89(2):479–499
- O'Sullivan D, Hakley M (2000) Agent-based models and individualism: is the world agent-based? *Environ Plan A* 32(8):1409–1425
- Parker DC, Berger T, Manson SM (2002) Agent-based models of land-use and land-cover change. *LUCC Report Series No. 6*, Indiana University
- Parker DC, Manson SM, Janssen MA, Hoffmann MJ, Deadman P (2003) Multi-agent system models for the simulation of land-use and land-cover change: a review. *Ann Assoc Am Geogr* 93(2):314–337
- Railsback SF, Lytinen SL, Jackson SK (2006) Agent-based simulation platforms: review and development recommendations. *Simulation* 82(9):609–623
- Read D (2010) Agent-based and multi-agent simulations: coming of age or in search of an identity? *Comput Math Org Theory* 16(4):329–347, Special Issue
- Sante I, Garcia AM, Miranda D, Crecente R (2010) Cellular automata models for the simulation of real-world urban processes: a review and analysis. *Landsc Urban Plan* 96(2):108–122
- Schelling T (1971) Dynamic models of segregation. *J Math Sociol* 1(2):143–186
- Silva EA (2010) Complexity and CA, and application to metropolitan areas. In: de Roo G, Silva EA (eds) *A planner's encounter with complexity*. Ashgate, Aldershot, pp 187–207
- Spencer GM (2012) Creative economies of scale: an agent-based model of creativity and agglomeration. *J Econ Geogr* 12(1):247–271
- Takeyama M, Couclelis H (1997) Map dynamics: integrating cellular automata and GIS through geo-algebra. *Int J Geogr Inf Sci* 11(1):73–91
- Torrens PM (2012) Moving agent pedestrians through space and time. *Ann Assoc Am Geogr* 102(1):35–66
- Torrens PM, Benenson I (2005) Geographic automata systems. *Int J Geogr Inf Sci* 19(4):385–412

- Torrens PT, O'Sullivan D (2001) Cellular automata and urban simulation: where do we go from here? *Environ Plan B Plan Des* 28(2):163–168
- Waldrop MM (1993) Complexity: the emerging science at the edge of order and chaos. Simon & Schuster, New York
- White R, Engelen G (1993) Cellular automata and fractal urban form: a cellular modeling approach to the evolution of urban land-use patterns. *Environ Plan A* 25(8):1175–1189
- Wolfram S (ed) (1986) Theory and applications of cellular automata. World Scientific, New York
- Wolfram S (2002) A new kind of science. Wolfram Media, Champaign, IL
- Wu F (1998) SimLand: a prototype to simulate land conversion through the integrated GIS and CA with AHP-derived transition rules. *Int J Geogr Inf Sci* 12(1):63–82
- Wu F, Webster CJ (1998) Simulation of land development through the integration of cellular automata and multi-criteria evaluation. *Environ Plan B* 25(1):103–126

Alison J. Heppenstall and Dianna M. Smith

Contents

63.1	Introduction	1235
63.2	Defining Spatial Microsimulation	1236
63.3	Static and Dynamic	1237
63.4	Microsimulation as a Tool for Policy	1238
63.5	Spatial Microsimulation Algorithms	1240
63.5.1	Deterministic Reweighting	1240
63.5.2	Conditional Probabilities	1241
63.5.3	Simulated Annealing	1243
63.6	Which Algorithm?	1244
63.7	Case Study: Estimating Smoking Prevalence Locally	1245
63.8	Conclusions	1250
	References	1251

Abstract

Spatial microsimulation is an excellent option to create estimated populations at a range of spatial scales where data may be otherwise unavailable. In this chapter, we outline three common methods of spatial microsimulation, identifying the relative strengths and weaknesses of each approach. We conclude with a worked example using deterministic reweighting to estimate tobacco smoking prevalence by neighborhood in London, UK. This illustrates how spatial microsimulation may be used to estimate not only populations but also behaviors and how this information may then be used to predict the outcomes of policy change at the local level.

A.J. Heppenstall (✉)

School of Geography, University of Leeds, Leeds, UK

e-mail: A.J.Heppenstall@leeds.ac.uk

D.M. Smith

Queen Mary University, London, UK

e-mail: d.smith@qmul.ac.uk

63.1 Introduction

Social science research increasingly aims to explore the effects of government policy on individual behavior. Although detailed data are available about individuals from disparate surveys and sources (such as health records), what is often not known/readily available is the spatial location of these individuals within a city or country. Microsimulation can link data from multiple sources to model behavior at the individual level, to identify groups within a population who may need additional support or assistance before changing as aspect of public policy. Microsimulation may also be applied spatially to estimate small-area population effects of public policy prior to implementation. In this way, governments may be better prepared for local changes that follow rollout of a new tax or urban planning scheme. Additionally, the knowledge of how particular behaviors vary spatially can lead to more efficient allocation of services: garbage collection in areas of highest waste generation and hospitals located in proximity to areas with more traffic accidents or greater prevalence of long-term illness.

The focus in this chapter is spatial, rather than individual, microsimulation. Spatial microsimulation creates a “synthetic” population (compiled from anonymized individual-level data) which realistically matches the population, as defined by the population census, in a geographical area for a given set of criteria (constraints). A diverse set of research and policy applications utilize spatial synthetic populations, including health (Brown and Harding 2002; Smith et al. 2011), transportation (Beckman et al. 1996), and water demand estimation (Williamson and Clarke 1996).

The accurate representation of populations within geographical areas may have great significance for microsimulation modeling when the outputs are intended to inform policy. The US and UK Censuses collect comprehensive sociodemographic data on individuals, but to protect confidentiality, data are aggregated to larger geographic scales. At a coarse geographic scale (such as US state or county), more attribute detail is available, including cross tabulation of attributes such as tables of age-sex-ethnicity distribution of the population. At a fine geographic scale (e.g., UK output areas with average populations of 250), detailed population attributes are not available, only univariate tables of age, sex, or ethnicity. The lack of detailed data at the local level has led to research focused on creating realistic synthetic populations within a predefined geographical area, estimating combinations of attributes and/or data not available within census datasets, effectively filling in the blanks.

Within this chapter, both microsimulation and spatial microsimulation are defined. Using the main application areas as examples, we highlight the strengths and weaknesses of the approach. The focus then moves onto the main algorithms that are used within spatial microsimulation: deterministic reweighting, conditional probabilities, and simulated annealing. An exemplar of the application of spatial microsimulation is provided through a case study involving the estimation of smoking prevalence in local areas within London, UK. Finally, a general discussion is presented offering suggestions for areas of future development.

63.2 Defining Spatial Microsimulation

Microsimulation is the generation at time $t = 0$ of a population sample P made up of n individuals $[P^1, P^2, \dots, P^n]$ where each individual, i , has a number of initial attributes $[a^i_1, a^i_2, \dots, a^i_m]$. The population is then updated to later times, t , so the attributes of individual i become functions of time $[a^i_1(t), a^i_2(t), \dots, a^i_m(t)]$. This allows the population to be used to model the effects of policy changes on individuals. While detailed statistics on each individual attribute are often available for the whole population (e.g., number of males/females, number of people in a given age band), information about the codependencies of attributes is not (e.g., how many people in the 20–30 age band are male/female), or there is no information on policy-relevant outcomes such as household expenditure on childcare, or individual-level health behavior such as physical activity or smoking. What is often available is a sample of the population which does contain all these attributes. In microsimulation, the initial population is often taken from population samples in large-scale surveys.

Spatial microsimulation is a type of microsimulation that recognizes the key role of geography in many of the processes being modeled. Each respondent from a survey dataset, which includes the attributes we want to estimate for the population, is given a probability (weight) to live in a specific location or spatial area (e.g., a ward), based on the known population structure of each area from the census. As discussed above, one of the key challenges for spatial microsimulation is the creation of a realistic initial population. Often population samples are only available at the coarse spatial resolution (e.g., at country/regional level) and not at the fine spatial scale required. At the fine scale, only statistics on the individual attributes are available. Spatial microsimulation techniques therefore need to generate a realistic population in each area, j .

Mathematically, the population of area j can be written as $P_j = [w_{ij} P^i]$ where the weight w_{ij} represents the number of people in the population of area j , characteristic i , with the attributes of person P^i from the sample or synthetically generated population. To ensure the proportions of each individual attribute are correct in the area, the w_{ij} need to be chosen so the mean absolute error (MAE) given by $\sum_{km} |T_j^{km} - T_j^{km}(obs)|$ is minimized for each area j . Here, T_j^{km} is the modeled number of people in area j with attribute k taking the value m , and $T_j^{km}(obs)$ is the equivalent number observed. The modeled value is given by $T_j^{km} = \sum_i w_{ij} \delta_i^{km}$ where δ_i^{km} takes the value of one if attribute k takes the value m for person i and zero otherwise. Ideally, the MAE would be zero for all areas, but in practice, this is not always attainable given just an initial sample of the whole population. Much of the focus of the rest of this chapter will be discussing methods for obtaining the weights w_{ij} .

63.3 Static and Dynamic

Microsimulation models are categorized as either static or dynamic. In static microsimulation, a large representative sample has rules (normally drawn from data analysis or the literature) applied to it to generate the synthetic demographic and economic characteristics expected at one point in time. Spatial population

simulations are focused on what the consequences of external information bring to the population; it does not model the changes in the population itself. The defining characteristic of static microsimulation is that there is no direct change of the individuals within the model during the simulation time period, instead we focus on adding attributes to the existing population dataset. A typical “what-if?” scenario would be: “If there had been no poll tax in 1991, which communities would have benefited most and which would have had to have paid more tax in other forms?” (See Ballas et al. 2005; Gilbert and Troitzsch 2005). Further examples of static microsimulation in this area are presented below in Table 63.1.

In dynamic microsimulation, individuals change their characteristics as a result of endogenous factors within the model; the populations update over time. Various degrees of direct interaction between micropopulation units can be found in dynamic microsimulations, for example, processes such as birth and marriage. These models rely on accurate knowledge of the individuals and the dynamics of such interactions. In a dynamic microsimulation, the updating of the dynamic structure is performed by “aging” the population through the application of transition probabilities (i.e., what is the probability of an individual getting married or having a baby?). The changes in the population itself are modeled with changes in an individual in 1 year having an effect on the characteristics in subsequent years. A typical future-oriented dynamic microsimulation scenario would be as follows: if the current government had raised income taxes in 1997, what would the redistributive effects have been between different socioeconomic groups and between central cities and their suburbs by 2011? (O'Donoghue 2001; Ballas et al. 2005; Gilbert and Troitzsch 2005).

Static and dynamic microsimulation each has benefits and drawbacks. Static models tend to be simpler programs than dynamic models, and because they are less computationally demanding, simulations can be run quickly. There is a general acceptance that dynamic models provide a more realistic long-term estimate of individual-level behavior (O'Donoghue 2001). However, the process of generating realistic behavior involves potentially unlimited interactions/interdependencies of the individuals when updating; this can result in dynamic microsimulations being computationally demanding (Ballas et al. 2005).

63.4 Microsimulation as a Tool for Policy

One of the most important advantages of microsimulation is that it enables us to examine the impact of policy changes on individuals even with sparse data. This distinguishes microsimulation from alternative methods, such as Bayesian estimation (Congdon 2006) (for details of general Bayesian methods, see Brunsdon, this volume) or multilevel modeling for small-area population estimation (Moon et al. 2007). These two approaches to small-area population estimation require cross-tabulated data at the spatial scale of the simulation output, limiting the scale of any simulations. The key advantage of microsimulation is that it has no such requirement and may be carried out from a series of univariate tables through an iterative process of recalculating weights.

Table 63.1 Name and description of examples of microsimulation models within the four main domain areas (Adapted from Birkin and Wu 2012)

Model name and domain	Origin	Description and example applications
(a) Tax benefits		
POLIMOD	UK	Demonstrates how VAT, National Insurance Contributions, and local taxes are calculated under different assumptions; entitlement to retirement pension and other non-means-tested social security benefits
STINMOD	Australia	Static microsimulation model of the tax and transfer systems. The rules of government programs are applied to individuals and aggregated to calculate outcomes for income units, families, or households
EUROMOD	Europe	Tax-benefit model that covers 15 countries. It provides estimates of the distributional impact of changes to personal tax and transfer policy at either the national or the European level
(b) Pensions		
PRISM	UK	Dynamic microsimulation of income from social security, earnings, assets, public and private occupational pensions, and retirement savings plans
SfB3	Germany	Analysis of pension reforms, the effect of shortening worker hours, distributional effects of education transfers, and interpersonal redistribution in the state pension system
DYNACAN	Canada	Projects the incidence, average levels, and variation in private pensions into the future as a function of birth year, age, and gender
(c) Health care		
PBS	Australia	Expenditure on pharmaceuticals by different types of households, resultant government outlays under the Pharmaceutical Benefits Scheme, and the remaining patient co-payment contributions
LifeMOD	UK	Model the lifetime impact of the welfare state through examination of health status over the life course and implications for health-care financing in the UK
LifePaths	Canada	A dynamic longitudinal microsimulation model of individuals and families which simulates the discrete events that together constitute an individual's life history
DRACULA	UK	Simulate response of traffic to different network layouts and control strategies; measure network performance from outputs of the average travel time, speed, queue length, fuel consumption, and pollutant emission
Paramics	US	Microscopic simulation of a range of real-world traffic and transportation problems handling scenarios ranging from a single intersection to a congested freeway or the modeling of an entire city's traffic system
VisSim	Germany	Models traffic flow in urban areas as a discrete, stochastic, time step-based microscopic model, with driver-vehicle-units as single entities. The model contains a psychophysical car following model for longitudinal vehicle movement and a rule-based algorithm for lateral movements (lane changing)

Table 63.2 Name and description of examples of spatial microsimulation models

Model	Origin	Description
SVERIGE	Sweden	Dynamic population model designed to study human eco-dynamics. Simulates spatial location and mobility of individuals. Developed for Sweden
SimBritain	UK	Dynamic simulation attempting to model British population at different geographical scales up to the year 2021
HYDRA	UK	Grid-enabled decision-making support system for health service provision
SMILE	UK	Dynamic spatial microsimulation model designed to analyze the impact of policy change and economic development on rural areas in Ireland

Spatial microsimulation has further advantage over other microsimulation models in its ability to explore spatial relationships and analysis of the spatial implications of policy scenarios. These advantages are reflected in the different applications where microsimulation has been applied. [Table 63.1](#) presents an overview of the main subject domains and models that microsimulation can be found within, while [Table 63.2](#) presents examples of spatial microsimulation models.

Many of the models in [Tables 63.1](#) and [63.2](#) share a common feature; they are concerned with the idea of “what-if?” simulations whereby the impact of new or alternative policy rules on the whole system or individual parts/components can be assessed. A simple example might be, “What would happen to the economic situation of local households if there is a change in child benefit tax?”

63.5 Spatial Microsimulation Algorithms

There are several established methods used for spatial microsimulation, specifically: deterministic reweighting ([Smith et al. 2011](#)), conditional probability (Monte Carlo simulation) ([Birkin and Clarke 1988](#)), and simulated annealing ([Openshaw 1995](#); [Voas and Williamson 2001](#)). These methods were selected due to their common application in geography (see [Voas and Williamson 2001](#); [Ballas et al. 2005](#)). Further details of each of these algorithms, including mathematical derivations, can be found in [Harland et al. \(2012\)](#).

63.5.1 Deterministic Reweighting

The deterministic reweighting algorithm was introduced by [Ballas et al. \(2005\)](#) and has been widely used in microsimulation models for health-care research. As described in [Smith et al. \(2011\)](#), the deterministic reweighting algorithm proportionally fits each individual record in the sample to the observed counts in each of the constraint tables (univariate tables which define the “known” population within an area based on a set of attributes) iteratively until each of the constraint variables have been included. Each iteration of reweighting first applies Eq. (63.1) to calculate an initial new weight which is subsequently

reweighted to reflect the observed population count in the subsequent constraint tables using Eq. (63.2) and then a scaling factor is included in Eq. (63.3):

$$nw_o^c = oldw^c \frac{tot_o^c}{tot_s^c} \quad (63.1)$$

$$NW_o^c = \sum nw_o^c \quad (63.2)$$

$$\frac{tot_o^c}{NW_o^c} nw_o^c \quad (63.3)$$

where nw_o denotes new weight for individual in area o ; NW_o total weight for all individuals in area o ; c is constraint subcategory, such as “male” in the gender constraint table; tot_o^c total population in constraint c in area o defined by a population census; tot_s^c total population in constraint c in survey data; and $oldw^c$ is initial starting weight for individual, based on predefined survey weights, or may be set to equal one.

For each zone, the sample is cloned with the initial starting weights and then reweighted using the procedure outlined above. Each time an individual is “cloned” into an area, the additional attributes associated with them beyond the constraints are also included. Each person in the survey is given a weight, or probability, of living in each area based on the census constraint tables for each area, and this weight adjusts as the algorithm cycles through the constraints. Through this process, a more extensive population profile is made available. The local population profile will include decimal values of the weights, so outcomes are expressed as proportions of the base population rather than individuals replicated with attributes. Perfect matching between the synthetic totals and sample totals from the reweighting algorithm cannot be reached for every zone. The more dissimilar the characteristics of a zone are from the distribution of characteristics in the sample, the greater the resulting error, as this method assumes all of the areas are relatively homogeneous. A zone with a high ethnic minority population will differ to the average distribution of ethnic minority groups in the sample, making such a zone less likely to match the constraint table perfectly. Therefore, despite this algorithm having no stochastic element and being completely deterministic, the order in which constraints are applied can produce different resulting populations as each new weight produced is a product of the weight calculated using the preceding constraint information. One way to adapt this algorithm so the estimates are more accurate is to group together similar areas and run them through the model together (Smith et al. 2011). This is illustrated in the example at the end of this chapter.

63.5.2 Conditional Probabilities

The conditional probabilities model is an adaptation of the synthetic estimation procedures first introduced by Birkin and Clarke (1988). It was originally designed to generate a synthetic population where no survey data

existed. With the availability of more survey information, the algorithm has evolved to execute equally well using a sample.

The algorithm initiates by creating a population with the characteristics of the first constraint and the associated probability calculated from the constraint table. For example, the first constraint is gender and the number of males and females in the first geographical zone are 120 males and 180 females. Therefore, 120 individuals are created with an associated characteristic of male, and a further 180 individuals are created with a characteristic of female.

The second constraint is marital status with three categories: married, single, and divorced/widowed. The probability of married, single, and divorced/widowed people appearing in the zone is first derived from the sample. It is simply the count of individuals in each category from the second constraint, divided by the total number of individual records appearing in the first constraint. If the sample contained 1,000 records of which 400 were male, of the males 160 were married, the joint probability of a male being married would be $160/400 = 0.4$. The remaining joint probabilities are calculated for the male sample and result in being male and married, $p(\text{male, married})$, is 0.4, $p(\text{male, single})$ is 0.4, and $p(\text{male, divorced/widowed})$ is 0.2. The male portion of the synthetic population is iterated through, with a random number greater than zero and less than or equal to one generated for each individual. If the random number is less than 0.4, a characteristic of married is added to that individual; if the random number is between 0.4 and 0.8, a characteristic of single is added to the individual; and finally, if the random number is greater than 0.8, a characteristic of divorced/widowed is added to the individual. This process is repeated for the female category from the first constraint.

Once all individuals in the current zone have been assigned the second constraint characteristics, the totals for each of the three categories are calculated and compared to the totals observed in the marital status constraint table. The initial starting probabilities are adjusted to represent the discrepancies between the observed constraint totals and those calculated from the synthetic population. For example, if 100 married people were expected in this zone but only 80 had been created, the initial male probability of being married, $p(\text{male, married})$, of 0.4 would be boosted using Eq. (63.4), as would the corresponding female probability, $p(\text{female, married})$:

$$np_c(x, y, z, \dots, n) = op_c(x, y, z, \dots, n) \frac{T_c^{\text{constraint}}}{T_c^{\text{synthetic}}} \quad (63.4)$$

where $np_c(x, y, z, \dots, n)$ is the new joint probability calculated and $op_c(x, y, z, \dots, n)$ the old or initial joint probability. $T_c^{\text{constraint}}$ is the total number of individuals in category c of the current constraint, and $T_c^{\text{synthetic}}$ the total number of individuals in category c of the synthetic population.

For the male population, Eq. (63.4) would become $0.5 = 0.4 \frac{100}{80}$ giving a new $p(\text{male, married})$ of 0.5. Each joint probability for the second constraint is adjusted

and the characteristic assignment process repeated. This process is iterated through until changes in the joint probabilities fall below a predefined threshold. Each constraint is added individually using this technique until all the required constraints have been incorporated, and then the next zone is calculated starting from the initialization of the synthetic population using the values from constraint one. Once this process is complete, the synthetic population can be saved if no additional attributes from the sample are required. However, if the researcher is interested in examining attributes from the sample not included in the constraints, usually because they are not available, then an additional final Monte Carlo sampling stage is required. For each individual in the synthetic population, the sample is entered at a random point and iterated through until the first record exactly matches the constraint attributes generated for the synthesized individual. When a matching record in the sample has been found, the extra attributes are copied to the synthetic record.

63.5.3 Simulated Annealing

As outlined by Davies (p. 6, 1987), “Simulated annealing is a stochastic computational technique derived from statistical mechanics for finding near globally-minimum-cost solutions to large optimisation problems.” The essence of the procedure is to start by creating a population as a random extract from the sample file, and by aggregating for the various constraints, the goodness of fit of the population to the constraining tables can be evaluated. From this population, an individual member is selected at random and replaced with another individual that is also selected at random from the sample. The aggregation and goodness of fit evaluation is repeated, and if the fit is improved, then the new individual replaces the old.

The feature which distinguishes simulated annealing, for example, in contrast to hill-climbing algorithms, is the incorporation of the Metropolis algorithm allowing both backward and forward steps to be taken when searching for an optimal solution (Otten and van Ginneken 1989). So even if the replacement leads to deterioration in the model fit, it will be allowed by the model as long as a certain threshold is exceeded. This threshold is often characterized as a “temperature” step – or annealing factor – as this method was originally conceived as a means to simulate the annealing process by which metals are cooled. As the algorithm proceeds, the (temperature) thresholds are reduced, and so, backward steps become progressively more unlikely, so that ultimately only climbing moves are permitted toward an optimized outcome.

Simulated annealing is similar to deterministic reweighting to the extent that weights are applied to members of a sample. However, in simulated annealing, these weights are zero or one representing selection or exclusion, whereas in deterministic reweighting, the weights are fractional. Simulated annealing is a heuristic hill-climbing algorithm rather than an iterative process (deterministic reweighting) or sequential estimation method (conditional probabilities). One of the

most important differences is that simulated annealing evaluates individual moves simultaneously against all of the constraining tables, whereas in both of the other techniques, this evaluation takes place constraint by constraint.

63.6 Which Algorithm?

Each spatial microsimulation algorithm/method possesses its own inherent advantages and disadvantages. The following section evaluates the strengths and weaknesses of each method in the context of issues to consider before selection of an algorithm.

First, how much preprocessing is required to obtain a robust fit-for-purpose output? As both the deterministic reweighting and conditional probabilities algorithms reweight the resulting populations using one constraint at a time, building on the results from the previous constraint, they are sensitive to the constraint order specified within the model; the first constraint will have the greatest impact on the final “weight” assigned to an individual to live in an area.

The primary difference between deterministic reweighting and conditional probabilities is the lack of a stochastic process in deterministic reweighting. As the deterministic process will give the same result every time, the impact of slight modification to the constraint variables will be clear from the results. To model the prevalence of type 2 diabetes using a deterministic method and estimate the impact of an aging population on diabetes prevalence, the age constraint may be adapted to reflect an aging population. The model is then rerun to estimate diabetes prevalence under this aged population distribution. In contrast, the simulated annealing algorithm places equal weight on each constraint (although this can be changed at the researcher’s discretion), and thus, the constraint order is of no consequence to the outputs.

The number of constraints that can be specified is related to the speed of execution of each algorithm. The deterministic reweighting algorithm reweights the sample using all of the constraint information in a specified order. As more constraints are added, the difference in sample and constraint frequency distributions can become more pronounced, especially at finer geographies where constraint populations are small. Therefore, less robust results may be produced as additional constraints are included in the model. The conditional probabilities model suffers with similar issues; however, these are less pronounced due to the joint probabilities for constraint combinations being adjusted in isolation. However, increasing the number of constraints increases both the processing time and the likelihood of being unable to converge on a suitable joint probability for a constraint combination. Simulated annealing also suffers from a time performance penalty as the number of constraints is increased although the rate increase is less severe than for the other two algorithms.

One of the major advantages that the conditional probabilities method has over the other two algorithms is that if a sample is unavailable, a synthetic population can be created using only the aggregate information from the

constraint tables. However, as the algorithm here requires a sample from which to extract the initial joint probabilities, an alternative source for this information is required in the absence of a sample otherwise erroneous individuals, such as married children, could be produced. A major advantage of the simulated annealing approach, as discussed above, is the inclusion of the Metropolis algorithm. However, the drawback to added search power is the associated higher computational times. Neither the conditional probabilities method nor the deterministic reweighting method can take backward steps when searching for a solution.

Finally, both the conditional probabilities and the simulated annealing methods contain a stochastic element that results in the creation of a different population configuration each time the model is run. This allows the model to consider and produce alternative and potentially more realistic populations. Deterministic reweighting does not have this capability. However, because deterministic reweighting produces the same result with each model run, the impacts of any starting constraint change are more easily quantifiable and can be important for policy evaluation, as discussed above.

There is only one example of these methods being compared and systematically evaluated within the published literature. Harland et al. (2012) compared the outputs of spatial microsimulation algorithms at varying spatial scales. While simulated annealing performed very well in each of the experiments that were performed, no clear winner was advocated. An interesting finding was that the simulation of attributes that are particularly influenced by spatial locations is best undertaken using the deterministic reweighting algorithm. This method allows the constraint order to be tailored to best represent a particular cluster of zones and is more accurate when the purpose is to model one distinct outcome rather than recreating a generic population (Smith et al. 2011). The following case study uses the deterministic reweighting algorithm to estimate one such outcome, estimation of smoking prevalence in London.

63.7 Case Study: Estimating Smoking Prevalence Locally

The application of spatial microsimulation to health outcomes is a valuable extension to the earlier models of tax policy and population estimation. Local-level public health data are more readily available due to electronic patient records and regular data collection as part of the funding scheme in the UK; however, these data only reflect individuals who are registered with a general practice or visit the hospital. Patient records are protected by strict information governance, and patient right to privacy or confidentiality prevents widespread access to or use of records which may inadvertently identify an individual. Within the UK data, governance prohibits most spatial mapping of individual health data below the Lower Super Output Area level [=LSOA] (about 1,500 individuals). Even when the data exists, it may not be used for spatial analysis if there is risk of reidentifying patients when their location is combined with demographic

characteristics. Small-area estimation of health outcomes/behaviors offers a solution to restricted spatial analysis and can be accomplished by combining detailed health surveys with population census data with spatial microsimulation.

Small-area estimation of health data may be carried out using a variety of statistical methods and frameworks. Here we focus on spatial microsimulation using a deterministic reweighting algorithm, although as mentioned previously, there are numerous alternatives including multilevel modeling (Moon et al. 2007) and Bayesian methods (Congdon 2006). In the example to follow, we will show that the microsimulation algorithm requires only univariate population data tables at the geographic level of interest. The population census restricts multivariate (cross-tabulated) tables at lower levels with the aim of protecting identity. For this reason, spatial microsimulation has been used to estimate health outcomes down to the output area level, which typically includes about 150 individuals. If we knew how smoking prevalence varied spatially, we would be able to allocate smoking cessation services and associated resources more effectively.

There are few complete population health censuses available beyond local bespoke surveys. More frequently, countries will have national-level health surveys (such as the Health Survey for England or New Zealand Health Survey) repeated either annually or every set number of years. These national surveys are conducted with a representative sample of the population and often include hundreds of variables to provide comprehensive profiles of respondents' health. As part of the data collection process, the surveys will also collect information on basic social and demographic characteristics of respondents. This data allows respondents from the health survey to be "linked" to people from the census who share the same demographic traits. This process will be illustrated in an example of smoking prevalence estimation among adults in London (2001 Census population: 7,172,090).

Data to produce the smoking estimates will come from the 2008 Health Survey for England [=HSE] ($n = 12,648$ respondents). In this example, we will use logistic regression to identify the social and demographic variables present in both the HSE and the 2001 Census that are best predictors of cigarette smoking among adults in England. There must be consistency in the variables between both datasets: the census provides the spatially defined population at the output area level, and the HSE gives us the probability of an individual who fits a given social and demographic profile to be a smoker. For the sake of simplicity, this example will be limited to four predictor variables (constraints).

The logistic regression model is run using SPSS 18.0. A series of potential predictor variables are identified from literature on smoking behavior: age, sex, ethnicity, social grade, marital status, employment status, and housing tenure. Smith et al. validated this method to predict smoking accurately in New Zealand (2011) using age, sex, ethnicity, and income data from the New Zealand Health Survey and the New Zealand Census of Population. After running a series of logistic regression models and comparing model fit, the four best predictors of smoking status from the HSE were identified as age, sex, marital status, and social grade. The constraints variables are categorical (Table 63.3).

Table 63.3 Constraints in the smoking model

2008 HSE data Constraint		<i>n</i>	%
Sex	Male	5,897	46.6
	Female	6,751	53.4
Age	0–17	3,069	24.3
	18–24	575	4.5
	25–34	1,096	8.7
	35–44	1,321	10.4
	45–54	1,305	10.3
	55–64	1,234	9.8
	65+	4,048	32
Social grade	AB	5,193	41.1
	C1C2	4,946	39.1
	DE	2,509	19.8
Marital status	Single	4,873	38.5
	Married	5,505	43.5
	Separated	211	1.7
	Divorced/widowed	2,059	16.3

NB The social groups in the UK are defined as follows: *A* is upper middle class (higher managerial, administrative, or professional), *B* is middle class (managerial, administrative, or professional), *C1* is lower middle class and *C2* skilled working class (junior managerial, skilled manual laborers), *D* is working class (semi- and unskilled workers), and *E* are those at the lowest level of subsistence (pensioners, widows)

There are a total of 4,765 LSOAs in London. The constraint tables are created with the total counts of individuals in each LSOA, which must sum to the same total population in each LSOA for each constraint variable. This must be checked carefully as some people will not answer all of the questions in the census. To ensure that there are the same total counts for each LSOA in each constraint, the numbers are proportionally adjusted to sum to the total provided in the basic population table for the LSOA. For example, in the first LSOA, the total population is known to be 1,600. However, only 1,540 people answered the marital status question. To adjust for the known total population (1,600), the number of people in group AB (670) is divided by the total count in the social grade table (1,540) then multiplied by 1,600:

$$670/1,540 \cdot 1,600 = 696.1$$

This is repeated for all of the categories in social grade, in each LSOA.

As discussed previously in the spatial microsimulation literature, with the deterministic reweighting algorithm applied here, the model will smooth all of the areas to look similar to each other. This will increase the error in prevalence estimates for areas which are unique, because the local population will look very different from the “general” population. One way to minimize the tendency to

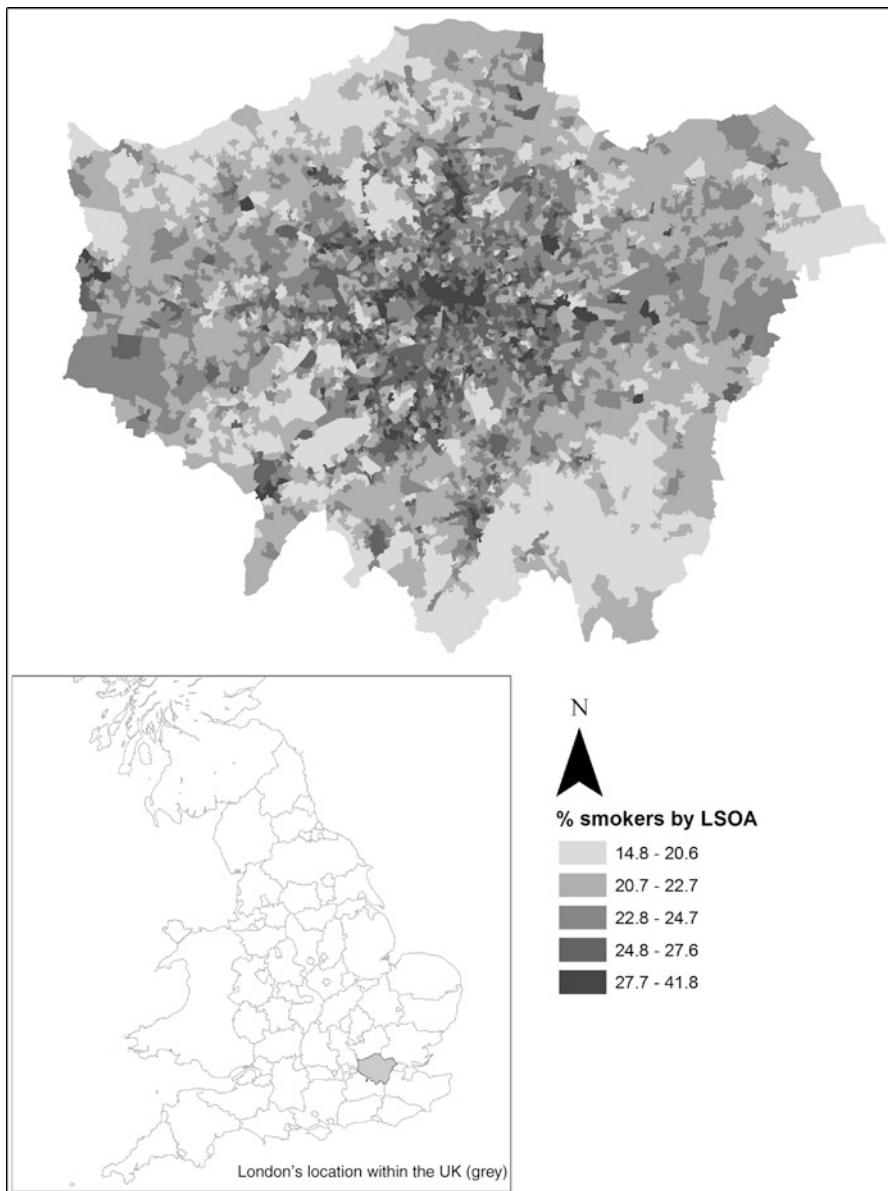


Fig. 63.1 Estimates of smoking prevalence of adults in LSOA's in London

higher levels of error in dissimilar places is to first identify the LSOAs which have a similar population profile in terms of the constraint variables used to predict the health behavior. This may be done by running a k-means cluster analysis (Smith et al. 2009). In this example, the LSOAs are clustered based on the percent of

Table 63.4 Error between microsimulation model and estimates from NHS information centre (2005–2008 data)

Primary care organization	Prevalence error (% simulation – % PCO model)
Barking and Dagenham	9.34
Barnet	-3.65
Bexley	6.03
Brent Teaching	-5.03
Bromley	0.94
Camden	-1.30
Croydon	1.45
Ealing	-4.38
Enfield	1.18
Greenwich teaching	3.64
City and Hackney teaching	1.27
Hammersmith and Fulham	-0.67
Haringey teaching	1.56
Harrow	-6.86
Havering	5.88
Hillingdon	0.18
Hounslow	-0.94
Islington	2.56
Kensington and Chelsea	-4.82
Kingston	-0.63
Lambeth	3.37
Lewisham	3.37
Newham	2.22
Redbridge	-4.20
Richmond and Twickenham	-1.61
Southwark	3.20
Sutton and Merton	1.26
Tower Hamlets	1.84
Waltham forest	1.14
Wandsworth	-0.49
Westminster	-2.59

the population in the 18–24 age range and social groups D or E and the percent who are unmarried. All of these groups are most likely to be smokers within the constraint data.

There are a total of five clusters based on these population characteristics. The model is then run five times, once for each of the clusters. The results are decimal values of people as smokers or nonsmokers, or children under the age of 18. The final counts of smokers are then used to calculate the prevalence of smoking among adults in each LSOA. The estimates are mapped at the LSOA level for London in Fig. 63.1. Estimated prevalence ranged from 14.8 % to 41.8 % by LSOA.

The results are validated against estimates created by the National Health Service (NHS: this is the main health-care provider in the UK) Information Centre (IC), based on HSE data from 2005 to 2008. The NHS estimates are available only at the primary care organization (PCO) level, of which there are 31 in London. We aggregated the estimates from the microsimulation model to the PCO level and measured the difference in prevalence between the NHS estimates and our estimates. The mean absolute difference was 2.85 % overall, with a range of values from -6.86 % to 9.34 % (Table 63.4). The greatest error was in Barking and Dagenham PCO (9.34 %). The error is reasonable given the available data and the time difference between our estimates and those created by the NHS IC.

The estimate error may also be measured against real-world data, where available (see Smith et al. 2011), or against a similar outcome such as diabetic amputations when predicting prevalence of diabetes (Congdon 2006). Alternatively, the error may be tested against another known value that is related to both the constraint variables and the outcome.

As this deterministic reweighting example shows, there are basic steps to the simulation process that must be conducted every time:

- Identify viable aspatial dataset for the health outcome/behavior (here, the HSE).
- Conduct statistical analyses to identify the optimal predictor variables for the outcome that are available from a spatial dataset (here, the census).
- Using the statistical information on predictors of the health outcome, cluster the areas based on the proportion of area populations that are in the highest subgroup of each constraint (i.e., Which social grade contains the greatest proportion of smokers? Marital status group?).
- Ensure that the constraint tables are prepared with the same count of people in each area across all variables.
- Run the model and validate against available data, similar outcome, or a related variable present in both the spatial and aspatial data.

63.8 Conclusions

Comprehensive socioeconomic data is not, for reasons of confidentiality, available at the individual level within any locality. This lack of detailed data has motivated research into spatial microsimulation with the intention of creating realistic synthetic populations that are representative of the geographical areas to which they represent. These populations can then input into simulation models that allow policy makers to understand the small-area impact in policy or demographics.

This chapter has provided a brief overview of spatial microsimulation, focusing on the main algorithms that are typically employed. A case study of prevalence of smoking in London was presented using deterministic reweighting to show the value that this approach can bring to informing policy on health outcomes. It is clear from this example that spatial microsimulation has a great deal to offer for social simulation modeling.

However, there remain many possible directions for future research. In terms of improving the modeling technique, there is no overarching consensus to which algorithm (reweighting/synthetic reconstruction technique) is the most appropriate or accurate for different domain applications or at which spatial level. This requires further research on validation of the synthetic population estimates produced by researchers. It also requires researchers to be honest about areas in which spatial microsimulation is not successful!

There are several criticisms that can be leveled at microsimulation. They are data hungry, computationally intensive, only model one-way interactions (the impact of policy on individuals), and weak in handling behavioral modeling. Several of these limitations can be overcome by hybridization with other individual-based models, in particular, agent-based models (see Crooks and Heppenstall 2012).

This is perhaps one of the most exciting areas of future research for spatial microsimulation. Realistic individual-level populations can be generated that mimic specific characteristics about a geographical area, or a specific application; for example, populations can be generated that contain characteristics of particular interest to health or education researcher. These populations can be turned into individual agents that form part of a larger modeling effort (see Wu and Birkin 2012). Hybridization allows the incorporation of both different types of behavior and detailed interactions between individuals, something which microsimulation alone is not capable of. The value of hybridization with agent-based models is that it would allow both researchers and policy makers to ask more sophisticated questions of social simulation models and in turn receive more accurate and realistic forecasts.

Acknowledgments This work was funded by the ESRC funded grant “Modeling Individual Consumer Behavior” (RES-061-25-0030) and MRC Population Health Scientist Fellowship (G0802447). The modeling framework used was developed by Kirk Harland.

References

- Anderson B (2007) Creating small-area income estimates: spatial microsimulation modeling. Department for Communities and Local Government. Communities and Local Government, London
- Ballas D, Rossiter D, Thomas B, Clarke G, Dorling D (2005) Geography matters. Simulating the local impacts of national social policies. Joseph Rowntree Foundation, York
- Beckman RJ, Baggerly KA, McKay MD (1996) Creating synthetic baseline populations. *Transport Res Part A* 30(6):415–429
- Birkin M, Clarke M (1988) SYNTHESIS – a synthetic spatial information system for urban and regional analysis: methods and examples. *Environ Plann A* 20:1645–1671
- Birkin M, Clarke M (1989) The generation of individual and household incomes at the small area level using SYNTHESIS. *Reg Stud* 23(6):535–548
- Birkin M, Wu B (2012) A review of microsimulation and hybrid agent-based models. In: Heppenstall AJ, Crooks AT, See LM, Batty M (eds) *Agent-based models of geographical systems*. Springer, Dordrecht, pp 51–68

- Brown L, Harding A (2002) Social modeling and public policy: application of microsimulation modeling in Australia. *Jasss J Artif Soc Soc Simul* 5:4
- Congdon P (2006) Estimating diabetes prevalence by small area in England. *J Pub Health* 28(1):71–81
- Crooks A, Heppenstall A (2012) Introduction to agent-based modeling. In: Heppenstall AJ, Crooks AT, See LM, Batty M (eds) *Agent-based models of geographical systems*. Springer, Dordrecht, pp 85–108
- Davies L (1987) *Genetic algorithms and simulated annealing: research notes in artificial intelligence*. Pitman, London
- Gilbert N, Troitzsch KG (2005) *Simulation for the social scientist*. Open University Press, Berkshire
- Harland K, Heppenstall AJ, Smith DM, Birkin MH (2012) Creating realistic synthetic populations at varying spatial scales: a comparative critique of population synthesis techniques. *J Artif Soc Soc Simul* 15:1
- Kennell DL, Sheils JF (1990) PRISM: dynamic simulation of pension and retirement income. In: Lewis GH, Michel RC (eds) *Microsimulation techniques for tax and transfer analysis*. The Urban Institute Press, Washington, DC
- Lambert S, Percival R, Schofield D, Paul S (1994) An introduction to STINMOD: a static microsimulation Model, NATSEM Technical Paper No 1. University of Canberra, Canberra
- Liu R (2005) The DRACULA dynamic network microsimulation model. In: Kitamura R, Kuwahara M (eds) *Simulation approaches in transportation analysis: recent advances and challenges*. Springer, pp. 23–56. ISBN0-387-24108-6
- Moon G, Quarendon G, Barnard S, Twigg L, Blyth B (2007) Fat nation: deciphering the distinctive geographies of obesity in England. *Soc Sci Med* 65(1):25–31
- O'Donoghue C (2001) Dynamic microsimulation: a methodological survey. *Brazilian Elect J Econ* 4:2
- Openshaw S (1995) Developing automated and smart spatial pattern exploration tools for geographical information systems applications. *Statistician* 44:3–16
- Openshaw S, Rao L (1995) Algorithms for reengineering 1991 census geography. *Environ Plann A* 27:425–446
- Otten RHJM, van Ginneken LPPP (1989) The annealing algorithm. *The Springer Int Ser Engin Comp Sci* 72(1):5–17
- Redmond G, Sutherland H, Wilson M (1998) The arithmetic of tax and social security reform: a user's guide to microsimulation: methods and analysis. Cambridge University Press, Cambridge
- Rephann TJ (1999) The education module for SVERIGE: Documentation V 1.0. Available at: <http://www.equotent.net/papers/educate.pdf>
- Smith DM, Clarke GP, Harland K (2009) Improving the synthetic data generation process in spatial microsimulation models. *Environ Plann A* 41(5):1251–1268
- Smith DM, Pearce JR, Harland K (2011) Can a deterministic spatial microsimulation model provide reliable small-area estimates of health behaviors? An example of smoking prevalence in New Zealand. *Health Place* 17:618–624
- Voas D, Williamson P (2000) An evaluation of the combinatorial optimisation approach to the creation of synthetic microdata. *Int J Popul Geogr* 6:349–366
- Voas D, Williamson P (2001) Evaluating goodness-of-fit measures for synthetic microdata. *Geograph Environ Model* 5:177–200
- Williamson P, Clarke GP (1996) Estimating small-area demands for water with the use of microsimulation. In: Clarke GP (ed) *Microsimulation for urban and regional policy analysis*. Pion, London, pp 117–148
- Williamson P, Birkin M, Rees P (1998) The estimation of population microdata by using data from small area statistics and samples of anonymised records. *Environ Plann A* 30:785–816
- Wu BM, Birkin MH (2012) Agent-based extensions to a spatial microsimulation model of demographic change. In Heppenstall AJ, Crooks AT, See LM, Batty M (eds) *Agent-based models of geographical systems*. Springer, Dordrecht, pp 347–360

David O'Sullivan

Contents

64.1	Introduction	1253
64.2	Spatial Networks and Graphs	1254
64.2.1	Basic Definitions	1255
64.2.2	Vertex Degree, Graph Density, and Local Clustering	1256
64.2.3	Spatial Embedding and Planarity	1258
64.2.4	Shortest Paths, Distances, and Network Efficiencies	1259
64.3	Higher-Order Structure in Networks	1261
64.3.1	Network Centrality	1262
64.3.2	Network Modules or Subgraphs	1262
64.3.3	Structural Equivalence	1263
64.4	Generating Networks: Spatial Network Models	1264
64.4.1	Spatial Networks From Point Patterns	1265
64.4.2	Spatial Small Worlds	1266
64.4.3	Growing Spatial Networks: Preferential Attachment	1267
64.4.4	Dual Graphs: New Graphs from Old	1268
64.4.5	Matrix and Adjacency List Representation of Graphs	1268
64.5	Properties of Real-World Spatial Networks	1270
64.5.1	Road Networks	1270
64.5.2	Transport Networks	1270
64.5.3	Other Spatially Embedded Networks	1271
64.6	Conclusions	1271
	References	1272

Abstract

Spatial networks organize and structure human social, economic, and cultural systems. The analysis of network structure depends on the development of

D. O'Sullivan

School of Environment, University of Auckland, Auckland, New Zealand

e-mail: d.osullivan@auckland.ac.nz

measures and models of networks, which in turn rely on mathematical graph theory. Key concepts and definitions from graph theory are reviewed and used to develop a variety of graph structural measures, which can be used to investigate local and global network structure. Particular emphasis is placed on high-level network structural features of centrality, cohesive subgraphs, and structural equivalence. Widely used models for spatial networks are introduced and discussed. Pointers to empirical research on real-world spatial networks are provided.

64.1 Introduction

It has become commonplace to think of ourselves as inhabitants of a “networked world.” The most obvious contemporary manifestation is the Internet, augmented in recent years by web 2.0 technologies that enable online social networks and by mobile technologies which maintain those connections even while people move through global transport networks from city to city and continent to continent. If “[t]he most profound technologies are those that disappear” (Weiser 1991, p. 94), then the Internet is by any measure profound, so much so that we only notice it – it only becomes visible – when it is unavailable. Of course, most networks are much older and more obviously geographical than the Internet. Significant infrastructure from transport systems and telecommunications to the supply of electricity and water is in the form of networks. Arguably, when it comes to understanding the aggregate geographies of the human world, whether from a social, economic, or cultural perspective, it is networks which structure, constitute, and organize those patterns.

Manuel Castells (1996) foresaw (but only just!) this development in his *The Rise of the Network Society*. Castells suggests that the network society alters social, economic, and cultural relationships, creating a global “space of flows” not directly associated with any particular location on the Earth’s surface. Less radically, other scholars have argued that a key determinant of the relative importance of world cities is not their geographical location *per se* but their location in economic, transport, social, and cultural networks. For example, Taylor et al. (2011), using network measures, rank the relative importance of cities to argue that London is an “alpha++” city outranking many more populous cities such as Tokyo (alpha+), Seoul (alpha), or Los Angeles (alpha). What makes London rank above other cities is not its particular individual characteristics or geographical location, but its position in relation to other cities, in other words its position in multiple overlapping networks of relationships between cities worldwide.

However, we are getting ahead of ourselves. Whether or not we consider a network analysis of world cities (or anything else) to be informative, before we can deploy such methods, we must define terms and develop measures. As in any field of quantitative study we need measures to enable repeatable descriptions of the objects of study and models to allow us to determine if the measurements we make of empirical cases are interesting. In the next section, basic concepts, definitions,

and measures from graph theory are introduced. There follows a consideration of higher-level concepts of graph structure and associated measures. Throughout these sections pertinent aspects of spatial networks are discussed. Following from this, we introduce some models for spatial networks and comment on their properties. We then consider some significant findings from the rapidly growing literature applying these methods to real spatial networks. The article ends with some pointers to possible future directions.

Note that we do not consider here the numerous problems in computer science, operations research, and transport analysis (particularly traffic assignment and related problems) which are closely associated with the analysis of spatial networks. Interested readers should consult reference works in these fields and related chapters in this major reference work.

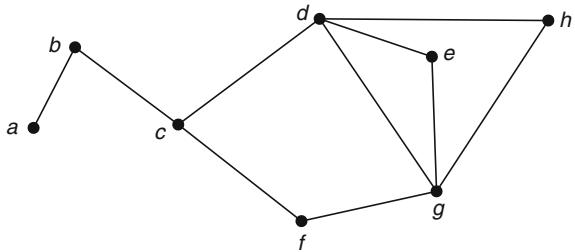
64.2 Spatial Networks and Graphs

In their stimulating and still relevant text *Network Analysis in Geography*, Haggett and Chorley (1969) follow Kansky (1963) in moving quickly from considering spatial networks to the analysis of mathematical graphs. Real spatial networks are complicated physical entities, with numerous elements, themselves often complex entities, such as multilane highways or airports with several runways (see Fischer 2004 for how these complications may be handled in a GIS setting). Our primary interest in the analysis of spatial networks lies in understanding how the network as a whole structures connectivity so as to centralize some locations, marginalize others, and, in general, differently position locations with respect to one another. It makes sense to strip away the messy complication of real spatial networks and work with the simpler, abstract representation of a mathematical graph.

We therefore begin with definitions from *graph theory*, which provides a foundation for the analysis of networks. A *graph* is a mathematical abstraction which can represent any set of elements somehow related to one another. Wilson (1996) provides a succinct introduction to the key terms and concepts discussed below. More advanced references delve into this field of discrete mathematics (Gross and Yellen 2006), which is fundamental to computer science (see, e.g., Jungnickel 1999), and is increasingly considered fundamental across all the sciences (Newman 2010).

64.2.1 Basic Definitions

A graph G consists of a finite, nonempty set $V = \{v_i\}$ of *vertices* and a finite nonempty set $E = \{e_i\}$ of distinct, unordered pairs of distinct elements in V , called *edges*. The number of elements in V , commonly denoted n , is the *degree* of G . The number of edges in E is often denoted m . Figure 64.1 shows a typical small graph with $V = \{a, b, c, d, e, f, g, h\}$, $E = \{ab, bc, cd, cf, de, dg, dh, eg, fg, gh\}$, $n = 8$, and $m = 10$.

Fig. 64.1 A typical graph

The edge $v_i v_j$, or e_{ij} , is said to *join* (more commonly link or connect) the vertices v_i and v_j , and these vertices are considered *adjacent*. We say that e_{ij} is *incident* with v_i and v_j and that v_i is a *neighbor* of v_j . The *neighborhood* $N(v_i)$ of v_i , often written simply N_i , is the set of vertices adjacent to v_i . Two edges incident with the same vertex are *adjacent edges*. In Fig. 64.1, $N_b = \{a, c\}$, and edges ab and bc are adjacent.

Given the ubiquity of graphs (or networks), it should come as no surprise that there is considerable confusion around terminology, with different fields adopting different terms in various contexts. Vertices are often referred to as *nodes* and represent the entities in a network, such as cities, people, cell phone towers, airports, and railroad stations. Edges are commonly referred to as *links* or *connections* and represent relationships between nodes, such as movements of goods, services or people, existence of airline routes, and mutual intervisibility. We can think of graphs as mathematical abstractions of networks which exist in the real world, in much the same way that variables represent measurements of real phenomena – this is the distinction between vertices and nodes, edges and links, and so on. In this section, while introducing formal definitions from graph theory, we adopt the proper mathematical terms, but elsewhere may return to widely used synonyms (such as network, node, and link).

The structure $G = (V, E)$ described so far is a *simple graph*, which has limited relevance to the representation of complicated real-world networks. We may also want to include cases where vertices may be joined to themselves by a *loop* $v_i v_i$, and multiple edges may also be allowed if we drop the requirement that edges be distinct. More significantly, *directed graphs* (sometimes referred to as *digraphs*) consist of a set of vertices V and a set of *arcs* A or *directed edges*, each of which consists of an *ordered* pair of vertices in V , implying directionality in the relationship between the vertices. This departure from the simple graph allows us to consider relationships where flows in each direction may be different (or even nonexistent in one direction), and is obviously an important consideration when we consider many real-world infrastructure or distribution networks.

Another variant on the simple graph is the *weighted graph* where each edge has an associated value or weight often denoted w_{ij} , representing some attribute of the relationship between the vertices it joins. The most obvious attribute of interest in many geographical applications is the length of the edge, measured either as a distance or perhaps duration. More generally, edge weights may represent some

cost associated with movement along the edge. Less obviously, but equally applicable, are edge weights that somehow represent the strength of the relationship between the vertices they join. The volume or value of trade between two countries and the number of flights daily between two airports are just two examples among many possibilities. In many cases, weights relating to the strength of a relationship between the incident vertices will reflect rates of flow or the capacity of the associated edges.

64.2.2 Vertex Degree, Graph Density, and Local Clustering

Even the limited graph theoretic concepts introduced so far allow us to develop useful descriptive measures of graph structure. Most obviously, the number of edges incident with a vertex is its *degree* denoted $\deg(v_i)$ or k_i . The average vertex degree is a useful summary measure of graph structure, given by

$$\bar{k} = 2m/n \quad (64.1)$$

since each edge is incident with two vertices. This measure is equivalent to Kansky's β index (1963) differing only by a constant multiplier. The *degree list* of a graph is the set of vertex degrees often arranged in order of increasing degree. For the graph in Fig. 64.1, the degree list is $\{1, 2, 2, 2, 2, 3, 4, 4\}$. In large graphs representing complex real-world networks, it is more useful to examine the *degree distribution* of the vertices, an aspect considered in more detail in later sections, although, as we shall see the degree distributions of many spatial networks are strongly constrained by their spatial embedding. If all vertices have the same degree k , it is *regular* of degree k , or k -*regular*. In practice, this is unlikely to occur in spatial networks, but may provide a useful benchmark or null model for assessment of how regularly structured is an observed network.

For a simple graph with n vertices, the maximum number of edges that could exist is given by $\binom{n}{2} = n(n - 1)/2$. Comparing the actual number of edges in the graph to this maximum provides a measure of how strongly connected the graph is overall, namely, its *density*, $\rho = m/\binom{n}{2} = 2m/n(n - 1)$. A graph's density is the fraction of all possible edges which *could* exist which actually *do* exist. The graph in Fig. 64.1 has density $2 \times 10/(8 \times 7) = 0.357$. Because the number of possible edges in a graph grows approximately with the square of its degree, whereas in most spatial networks the number of edges grows roughly linearly with graph degree, most spatial networks have low density, and only a small proportion of all the possible connections exist. This generally arises either because of distance decay effects or due to planarity constraints. We consider both issues in more detail below.

Because of the constraints on overall graph density in spatial networks, it is often more interesting to consider the local density or clustering of a spatial network.

This is a measure of how strongly connected the graph is in the neighborhood of each vertex. The *clustering coefficient* of a particular vertex is given by

$$C(v_i) = \frac{2m_i}{k_i(k_i - 1)} \quad (64.2)$$

where m_i is the number of edges joining vertices in the neighborhood of v_i . This is a direct localized equivalent to graph density and provides information about how well connected the network is locally. The distribution of the clustering coefficient in a network provides useful information about its structure. One interpretation is that it gives the probability, given that two vertices v_j and v_k are neighbors of v_i that v_j and v_k are themselves neighbors. Many spatial networks exhibit high clustering coefficients compared to nonspatial networks, which is unsurprising: if two vertices in a spatial network are neighbors, it implies that they are near one another, and if two vertices share a common neighbor, since they are probably also near one another, there is a high chance that they will be neighbors of one another.

64.2.3 Spatial Embedding and Planarity

Thus far, there has been no explicit consideration of the spatial aspect. Where we are concerned with spatial networks, vertices will have an associated spatial entity, often conveniently considered to be a point location, but potentially also a more complex spatial entity such as a region – for example, in a trade network, vertices may represent regions or countries.

Two types of spatial embedding of a graph are possible. The most obvious spatial networks are those where both vertices and edges are spatially embedded. Examples include transport and infrastructure networks, where the graph edges are physically realized in space, with direct implications for any associated weights or directional restrictions. Less obviously, spatial embedding of edges imposes a constraint on the overall network structure, that of *planarity*. A planar graph is one which can be drawn in two dimensions with no edges intersecting except at vertices on which they are both incident. For many infrastructure networks, this is approximately true, although bridges and tunnels in ground-transport networks are an obvious (but generally minor) exception. The planarity constraint significantly alters the overall structure of graphs, and we consider its implications in the following paragraphs.

A second form of spatial embedding is where vertices have associated spatial locations, but edges represent nonspatial relationships. An example is a spatially embedded social network. Individuals in the network have some spatial location – perhaps their home address – but edges might represent friendship or acquaintanceship relationships with no corresponding physical realization. A less obvious example is when the vertices in the graph represent spatially extended entities – such as metro lines – and edges represent a relationship such as “has an

interchange with.” Such networks rarely constitute the primary object of analysis, although they may easily arise as *dual graphs* in some analyses. In considering such a network to be a spatial network, we implicitly assume that the distance between vertices (whether direct Euclidean distance or over intervening spatial networks – see below) has an effect on the probability of their existence. In other words, we expect that vertices more remote from one another are less likely to be joined than those that are closer together.

Where the distinction matters, we will refer below to fully embedded or vertex-embedded spatial networks, reserving the term spatial networks to refer to networks of either kind.

The fundamental difference between spatial networks with spatially embedded edges, which are (approximately) planar, and spatial networks not affected by this constraint lies in the limits it places on the overall density of the graph both globally and locally. A fundamental result is Euler’s formula for planar graphs

$$n - m + f = 2 \quad (64.3)$$

where n and m are the number of vertices and edges as before, and f is the number of *faces* or regions in the plane which the graph divides the space into. We consider the overall region in which the graph is embedded as face, so that for the graph in Fig. 64.1, $f = 4$, that is, the whole space and the regions *cdgf*, *deg*, and *dhge*. Euler’s result is easily proved when we consider starting from a graph consisting of one vertex and no edges, so that $n = 1$, $m = 0$, and $f = 1$ when Eq. (64.3) clearly holds (see Fig. 64.2). Adding any edge while maintaining planarity, either (i) joins two existing vertices without intersecting an existing edge, so increasing both m and f by one, while leaving n unchanged, or (ii) adds a new vertex and joins it with an edge, increasing both n and m by one with no change in f . In either case, Eq. (64.3) remains true.

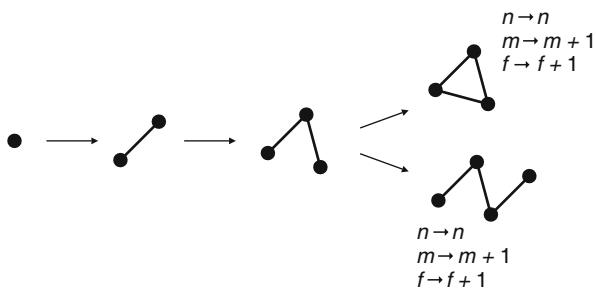
Euler’s formula has important implications for the possible density of planar graphs. Since every face requires at least three edges, and each face can share an edge with at most one other face, we know that $m \geq 3f/2$. Combining this result with Eq. (64.3) we arrive at

$$m \leq 3n - 6 \quad (64.4)$$

Combining this result with Eq. (64.1) tells us that the upper bound on the mean degree of a planar graph is $\bar{k} \leq 6$. Kansky (1963, p. 18) recognizes this in providing alternative formulations of his γ index (equivalent to graph density) for planar and nonplanar graphs.

Understanding this result, it is much easier to understand why area maps are so distinctively structured and why the Voronoi tessellation and associated Delaunay triangulation exhibit such characteristic structure. Since the spatial network constructed from the adjacency relations of a set of polygonal regions must necessarily be planar, the mean number of neighbors of each region cannot exceed 6. In graph terms, this bound on the number of edges in a planar graph

Fig. 64.2 How a planar graph grows as edges are added. Either the number of faces f (upper path) or the number of vertices n (lower path) must increase, but not both



and thus on many spatial networks, means that almost all spatially embedded networks are *sparse*, with $m \ll n^2$ and $\rho \propto 1/n$ as $n \rightarrow \infty$. In terms of local clustering, planarity also implies that any vertex with $k_i > 4$ must have $C_i < 1$ since it is impossible for any graph of more than four vertices to be fully connected.

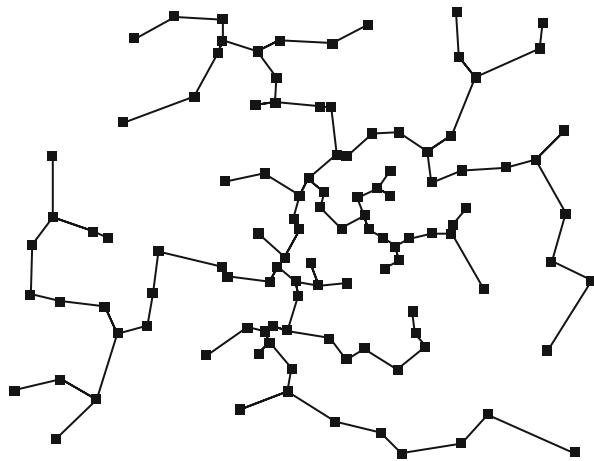
64.2.4 Shortest Paths, Distances, and Network Efficiencies

A particular sequence of edges $\{v_0v_1, v_1v_2, \dots, v_{r-1}v_r\}$ forms a *walk* of *length* r . If the vertices in a walk are distinct, then it is a *path*, and a path that begins and ends at the same vertex forms a *cycle*. The *distance* between vertices v_i and v_j , d_{ij} is the length of the shortest path among the set of all possible paths between v_i and v_j . Any one path between v_i and v_j of length d_{ij} is a *geodesic*. The largest distance between any two vertices is the *diameter* of the graph.

In a directed graph, walks may only proceed in the direction of constituent arcs. In a weighted digraph, the length of a path is generalized from the above definitions by summing the weight of its constituent edges, and the distance between two vertices is the length of the shortest path, as before. Given that the weights associated with the arcs in each direction between any two vertices are not necessarily the same (that is, $w_{ij} \neq w_{ji}$), there is no guarantee that the distances between vertices in a directed graph will be symmetric. Note also that where graph weights do not represent a “traversal cost,” such as when they represent link capacities or trade volumes, numbers of people, or other similar measures, then it does not make sense to accumulate edge weights in this way.

The (graph) distances between any two nodes or between all pairs of nodes in a spatial network are of considerable interest, particularly in how they compare to the corresponding straight line (Euclidean) distances between the corresponding node locations. If a network provides a path between two nodes whose distance is close to the straight line distance, then the network is efficient for that particular journey. On the other hand, if the network requires a much longer and more circuitous path to be taken between two locations, it is inefficient. A measure of the network efficiency for a single particular path is the *route factor* defined by Black (see 2003), following Nordbeck (1964) as

Fig. 64.3 The minimum spanning tree of a set of points



$$Q_{ij} = \frac{d_G(v_i, v_j)}{d_E(v_i, v_j)} \quad (64.5)$$

where d_G and d_E are graph-based and Euclidean distances, respectively, between locations i and j . We can average this quantity over a particular node

$$\bar{Q}_i = \frac{1}{n} \sum_j Q_{ij} \quad (64.6)$$

or over the whole network

$$\bar{Q} = \frac{1}{n(n-1)} \sum_{j \neq i} \bar{Q}_{ij} \quad (64.7)$$

The route factor provides one perspective on the efficiency of a network as-built. Another perspective is to consider how cheaply a set of locations might be connected. A *tree* is a graph which includes no cycles, in which $n = m + 1$. The *minimum spanning tree* of a set of vertices is the tree which minimizes the total weight of the edges in the tree, and when the weights relate to the cost of providing the associated node-to-node links, it represents the cheapest way to connect every node to every other. However, such a network is unlikely to be efficient from the point of view of a user of the network as measured using the route factor, since it will certainly involve many very circuitous shortest paths (see Fig. 64.3). Such a network is also vulnerable to failure, since losing just one edge will leave it disconnected. In practice, real networks will have more edges than the minimum spanning tree.

64.3 Higher-Order Structure in Networks

The measures we have considered so far generally focus on the overall structure of a network in a general way or on the structure at a particular location. Summary measures or distributional properties of these measures are generally useful, but they often fail to reveal structural aspects of networks which arise out of the totality of all the spatial relationships in the network. In this section, we briefly consider measures of such higher-order structure.

64.3.1 Network Centrality

Consideration of distance in networks leads naturally to questions of the most accessible or central locations in the network. An obvious approach is to calculate the mean distance from a node to every other node in the network:

$$\bar{d}_i = \frac{1}{n} \sum_j d_{ij} \quad (64.8)$$

where d_{ij} is the graph distance between vertices v_i and v_j as previously defined. Using this *centrality* measure, the most central node in a network is that with the minimum d_i . While this is an obvious measure of network centrality, there are many alternatives. One which has received considerable attention in recent years, because of its close relationship to movement on the network and to how subregions of the graph are connected to one another (see below) is *betweenness centrality*. The betweenness centrality of a vertex v_i is the proportion of the shortest paths between all other pairs of vertices $v_j \neq v_k$ in which v_i appears. If $g_{jk}(v_i)$ is the number of shortest paths from v_j to v_k in which v_i appears, and g_{jk} the total number of shortest paths from v_j to v_k , then

$$c_{between} = \sum_{j \neq k} \frac{g_{jk}(v_i)}{g_{jk}} \quad (64.9)$$

This measure has the nice property that it can be readily extended to edges also, simply being the proportion of all shortest paths on which each edge lies. Betweenness centrality provides an indication of the extent to which each vertex or edge has the potential to control movement or communication in the graph, assuming that there is “everywhere-to-everywhere” movement in the system. This measure is directly related to approaches that rely on random walk models. The most central vertices and edges measured in this way are those which will experience the most traffic when a population of random walkers move around the system.

64.3.2 Network Modules or Subgraphs

A class of measures which remains difficult to define precisely, but which has a clear intuitive interpretation, has recently come to the fore, as researchers attempt to determine how a network can be broken into cohesive subgraphs or regions, now most often referred to as *communities*. Fortunato (2010) provides a comprehensive overview of developments in this field. The general definition of a community is that the member vertices of a community are more strongly connected to one another within the community, than they are to vertices outside the community. This definition is not very precise, however. To take it to the extreme, we could argue that any joined pair of vertices are more closely connected to one another than they are on average to the rest of the graph (unless the graph is fully connected). A less extreme definition is to consider as communities, small, fully connected subgraphs or *cliques* (from their origins in social network analysis, see Wassermann and Faust 1994), but due to spatial constraints, cliques are unlikely in fully embedded spatial networks and so unlikely to be useful.

We obviously need a more flexible definition. Many have been suggested in the social networks literature (see Wassermann and Faust 1994, pp. 257–267), but most suffer from serious computational challenges in identifying them in graphs of any size, because of the exponential growth in the number of subsets of the vertex set V as graphs get larger. An important breakthrough has been in the development of more computationally tractable methods, beginning with the Girvan-Newman algorithm (Girvan and Newman 2002). Many of these methods are based on heuristic approaches, which successively remove edges of low betweenness centrality while repeatedly recalculating some measure of the quality of the resulting graph decomposition. Other methods aim to identify hierarchically nested communities and can consequently deal with very large networks with millions of nodes. Fortunato (2010) provides comprehensive details and references. It is notable that none of these methods are explicitly spatial, although where edge existence and/or weight is dependent on spatial proximity, this should not be a cause for concern.

The net result of these considerations is that the current working definition of a graph community is a circular one: graph communities are those subgraphs in a graph identified by a community-detection algorithm. This places considerable importance on the analyst's ability to meaningfully interpret any communities so identified, a situation analogous with that in the cluster analysis of multivariate statistical data.

Vertex centrality and community structure for a typical spatial network are illustrated in Fig. 64.4. In Fig. 64.4a vertex centrality calculated from the total path length from each vertex to every other while considering the Euclidean length of the graph edges is shown, with the darkest shaded vertices the most central. As is often the case, the most central vertices are those that are most geographically central to the network, as we might expect. By contrast, in Fig. 64.4b the betweenness centrality based only on the network topology is shown. Because, in topological terms, the vertices to the west of the central part of this network provide a shortcut around the densely packed central region, these vertices are highlighted

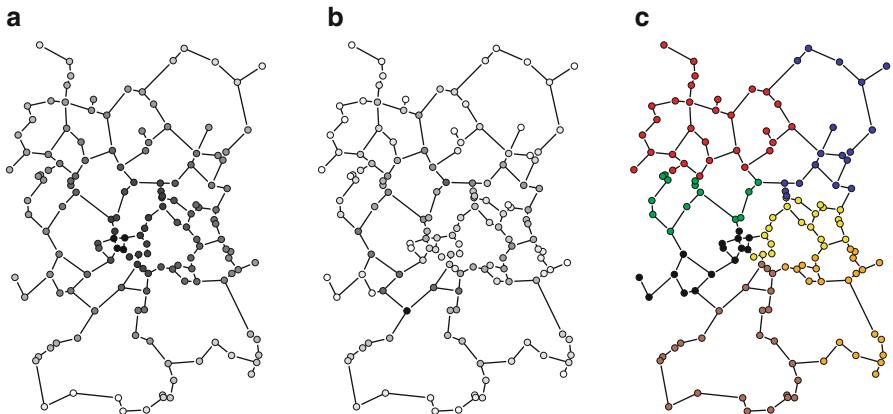


Fig. 64.4 Centrality and network communities illustrated. See text for details

by this centrality measure. Finally, Fig. 64.4c shows a possible community structure in this network where seven distinct regions in the network have been identified based on their mutual connectivity.

64.3.3 Structural Equivalence

A final graph structural characteristic likely to be of interest is the concept of *structural equivalence* among vertices or edges. This concept is easily grasped, but precise definitions are mathematically challenging and detection of structurally equivalent sets of vertices remains difficult. The idea is that vertices that are structurally equivalent have similar relationships to the rest of the graph as one another, an idea whose origins lie in social network analysis where structural equivalence is related to social roles (Lorrain and White 1971). Graph communities are a special case of structurally equivalent vertices, which share the property of being in the same community. In a transport network, we might expect major junctions on arterial routes to constitute an equivalence class. However, pinning down how this concept can be realized in practice has proved difficult, and detection of structurally equivalent (or more usefully structurally similar) sets of vertices is computationally challenging – consider that while community detection can assume that the subgraphs of interest are connected subsets of the graph, no such assumption can be made for structural equivalence classes. While the concept of structural equivalence is an attractive one for the analysis of spatial networks, progress in this area remains rather limited.

64.4 Generating Networks: Spatial Network Models

While measures of network structure are important tools in improving our understanding of spatial networks, it is equally important to develop models for

network formation. This was recognized by Haggett and Chorley in their coverage of [network] “Growth and Transformation” (1969, pp. 261–318), an extensive chapter and a very modern treatment. A recent review paper (Barthélemy 2011) provides a useful overview of many different spatial network models. Here we briefly review some of the available models and consider their general properties.

An important null model for any network is the Erdős–Rényi (E–R) model (see Erdős and Rényi 1960), which has been much studied. The E–R graph is generated as follows: create a set of n vertices, then consider every possible pair of vertices, and with probability p join them with an edge. Many of the expected properties of E–R graphs are well known. Of particular interest are the expected mean clustering coefficient and mean path length of vertices in the graph, once the network is sufficiently dense to be connected with no isolated clusters, an event which happens quite suddenly close to $p = \ln n/n$. The expected clustering coefficient $\langle C \rangle$ is given by p since p is the probability that any two vertices will be connected, and so is also the likely proportion of the neighbors of any vertex that will be connected. The expected shortest path length $\langle d \rangle$ in the E–R graph is approximated by $\ln n / \ln \bar{k}$.

64.4.1 Spatial Networks From Point Patterns

Perhaps the most obvious way to generate a spatial network is to begin with a point pattern and then to apply some geometric rules by which points are connected to one another (or not). This *geometric graph* model admits considerable variety in the outcomes depending on both the underlying point pattern and on the geometric rules applied. It also has the property that if we make the “rule” for joining nodes independent of the distance between them, then it is equivalent to the E–R random graph. More reasonable rules will be familiar from the construction of spatial weights matrices (see, e.g., pages 200–205 in O’Sullivan and Unwin 2010).

A *distance criterion* where two nodes are joined if they are closer together than some threshold distance. In Fig. 64.5a nodes nearer than 5 units apart are connected.

A *nearest neighbor criterion* where each node is joined to its nearest k neighbors. In Fig. 64.5b each node is joined to its 4 nearest neighbors.

An *attribute-distance rule* where depending on some attribute of the nodes and their separation distance they are joined or not. The simplest form of this rule is where the attribute is a radius of influence r_i , and nodes i and j are joined if $r_i + r_j < R$ where R is a threshold distance. An example is shown in Fig. 64.5c. Such a simple rule might be meaningful in the context of trees in a forest influencing one another, but in regional science, a more likely formulation will be based on an interaction measure such as $m_i m_j d_{ij}^{-\alpha}$ where the m values represent activity or population at each location and α is a constant controlling the rate at which likely connection falls away with distance.

A *pure geometric rule* such as those governing the Delaunay triangulation or closely related Gabriel graphs (see Okabe et al. 2000) shown in Fig. 64.5d, e, respectively.

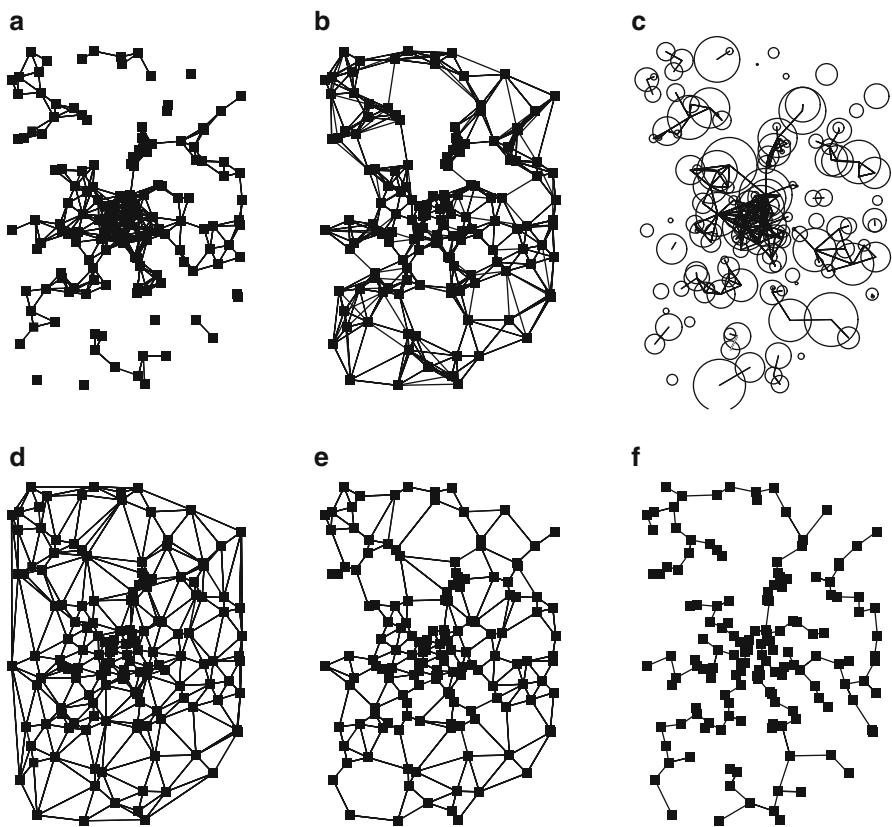


Fig. 64.5 Examples of geometric networks as described in the text. The region is 40 units east–west and 60 north–south, and the point pattern is inhomogeneous Poisson with greater intensity at the center of the region. Point symbols in (c) are scaled so that if the circles of two points intersect, they are joined in the network

A *global rule* such as that governing construction of the minimum spanning tree, where the edges are those which together connect the network with the minimum total path length. An example is shown in Fig. 64.5f.

As is clear from Fig. 64.5, the various geometric rules produce quite different networks. An important distinction is that the Delaunay, Gabriel, and minimum spanning trees are planar, whereas the other networks are not. At the same time that they do no guarantee planarity, the distance and distance-attribute models may also leave some nodes unconnected. It is unusual for real-world spatial networks to leave isolated regions, and so it may be that hybrid models where a distance criterion is applied subject to a connectivity and/or planarity requirement are more reasonable in some cases. One approach is to start with a network substrate, such as the minimum spanning tree or Gabriel graph, and add additional edges according to a distance criterion.

64.4.2 Spatial Small Worlds

Small-world networks are so-called after the commonly encountered apparent contradiction in social networks that while they are locally highly connected (i.e., they have high clustering coefficients), they also have globally short paths (i.e., the mean path length is low). It is apparent that this result does not hold for graphs generated by the E–R model. As noted, E–R graphs become connected when $p = \ln n/n$ when $\bar{k} \simeq \ln n$. For a network with 1,000 nodes, this gives us $\langle C \rangle = p = 0.00691$, $\bar{k} \simeq 6.91$, and $\langle d \rangle \simeq 3.57$. Increasing n to 10^6 reduces $\langle C \rangle$ to 1.38×10^{-5} while $\langle d \rangle$ only increases to 5.26. Clearly, although E–R networks are small worlds with short path lengths, they do not have the high local connectivity that makes this property in social networks surprising.

Watts and Strogatz (1998) presented an alternative network model that is both highly clustered locally, yet has short mean path lengths. Their approach is to start with a regular lattice and “rewire” it by breaking links and reconnecting them to other nodes selected at random from anywhere in the network. They show that only small numbers of rewiring events are necessary to dramatically reduce the mean path length in a lattice. Although Watts and Strogatz present their work for one-dimensional lattices, the basic idea is readily extended to more realistic spatial settings.

In two dimensions, a regular lattice is a grid of nodes with each node connected to its four nearest neighbors. The expected path length between any two nodes selected at random scales with $n^{1/2}$, and, in general in a D -dimensional lattice path lengths will scale with $n^{1/D}$. *Spatial small-world* models, rather than rewire the lattice, typically introduce additional “shortcut” links with the probability of the shortcuts dependent on the distance between the vertices they join. The probability that a shortcut e_{ij} exists might be proportional to $d_{ij}^{-\delta}$ where δ is a parameter chosen in a particular case. As δ is increased while holding constant the overall number of additional links added, the networks produced by models of this kind transition from random to small-world to regular lattice properties. This is readily understood in qualitative terms. For low values of δ , any length shortcut is equally likely – in effect, nodes are undifferentiated from one another – and the network has distance properties similar to a random network. High values of δ heavily penalize the provision of longer shortcuts, leaving the lattice’s overall $n^{1/D}$ distance-scaling property intact. Many transportation networks lie somewhere on this continuum, depending on how we incorporate shortcuts such as urban orbital highways, high-speed rail links, and airline routes into the more densely connected local transport network.

64.4.3 Growing Spatial Networks: Preferential Attachment

The examples above apply a connection or rewiring rule to a preexisting set of nodes. Arguably, a more realistic approach is to grow a spatial network from an initial individual node, by progressive addition of new nodes and edges, according to some rules governing how new nodes are attached to existing ones. Once again, the baseline case is a nonspatial network growth model known as the *preferential*

attachment model, attributable to Albert et al. (1999), which has spawned a large literature on “scale-free” networks (see Caldarelli 2007). The basic idea is that nodes are added to a network and attach themselves preferentially to those nodes that already have larger numbers of connections. The resulting networks have heavy-tailed distributions of vertex degrees meaning that a small number of very strongly connected nodes dominate the network structure.

Planar networks clearly cannot exhibit such characteristics, and physical constraints in most spatial networks prevent an unrestricted preference for attachment to the most well-connected existing nodes. Preferential attachment models that consider space, require each new node to have a spatial location, and the probability of attachment to existing nodes is then a function of both the degree of existing nodes and the distance between the new node and the existing nodes; this is similar to the attribute-distance geometric models considered previously, but with progressive addition of nodes rather than an all-at-once calculation. Models of this general structure often produce the hub-and-spoke structures characteristic of many distribution networks. Again, as with spatial small-world networks, the rate at which the probability of a connection decays with distance is important in determining overall characteristics of the resulting networks.

64.4.4 Dual Graphs: New Graphs from Old

An important idea for models of networks is the dual transformation, whereby an initial graph is transformed to a new graph by switching between nodes and edges or (in a planar graph) between faces and nodes. The *line graph* $L(G)$ of G is the graph whose vertices correspond to the edges of G and where two vertices are joined when their corresponding edges are adjacent. This dual transformation is shown in Fig. 64.6 and as in the case illustrated results in a denser graph with more variety in the vertex degree distribution than the original “primal” graph. The line graph dual transformation is often applied to the more obvious primal network representation of a system, such as the road intersection and segment network, because the richer structure provides more opportunities for insight into key features of the network. Figure 64.6b–d shows a simple example. In a planar graph, a similar dual transformation entails treating each face of the graph as vertex in a new graph, and joining those vertices whose faces are adjacent in the original graph. This is the relationship between the familiar Voronoi tessellation and the Delaunay triangulation (see Okabe et al. 2000).

64.4.5 Matrix and Adjacency List Representation of Graphs

Before closing the discussion of network analysis measures and models, it is important to note that even simple analysis of graphs requires careful consideration of how they are stored for computational purposes. There are two distinct approaches, which is preferable being largely a function of the graph density.

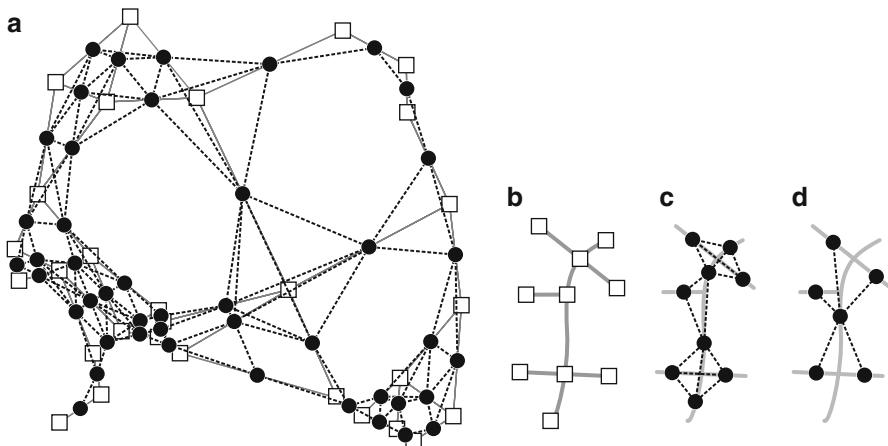


Fig. 64.6 Different graphs from the same network: (a) the line graph dual transformation, *white squares* and *gray lines* are the original graph and *black circles* and *dashed lines* are the line graph; (b) a primal graph for a road network; (c) line graph from the same road network; and (d) named road graph from the same road network

An obvious approach, given its close relationship to spatial weights matrices, is the graph *adjacency matrix* $\mathbf{A}(G) = a_{ij}$ where $a_{ij} = 1$ if the edge e_{ij} exists and 0 otherwise. The row order of \mathbf{A} is unimportant, but the row and column ordering must be the same. The *incidence matrix* \mathbf{B} is an alternative matrix representation that records the incidence of edges and vertices and is an $n \times m$ matrix where $b_{ij} = 1$ if e_j is incident with v_i , and 0 otherwise. A useful relationship between the incidence matrix and adjacency matrix is that

$$\mathbf{B}\mathbf{B}^T = \mathbf{A}^* \quad (64.10)$$

where \mathbf{A}^* is the adjacency matrix, modified such that the elements in the main diagonal are equal to the degree of the corresponding vertex. Another useful transformation is that the adjacency matrix $\mathbf{A}(L)$ of the line graph of G is given by $\mathbf{B}^T\mathbf{B} - 2\mathbf{I}_m$ where \mathbf{B} is the incidence matrix of G as defined above and \mathbf{I}_m is the $m \times m$ identity matrix.

However, many spatial networks have very low densities. This makes adjacency matrices an inefficient representation because many 0 entries are stored even though they record no useful information. Therefore, an *adjacency list* representation is often more appropriate and simply consists of a list of all the edges in the graph. Depending on the implementation details, it may also be necessary for vertices to be explicitly listed, or for the number of vertices in the graph to be stored, before the edges are listed. Appropriate modifications of such data structures can readily accommodate directed or weighted graphs.

Many tools and software platforms used for the analysis of graph data make use of both matrix and adjacency list representations, and, given the sparseness of many

spatial networks, it is important that an efficient *sparse matrix* implementation be available. For many analyses, the ability to quickly convert back and forth between (dense or sparse) matrix and adjacency list representations is necessary for efficient analysis.

64.5 Properties of Real-World Spatial Networks

Armed with the measures and models introduced above, it is possible to investigate the properties of real-world spatial networks. This remains an active area of research in many fields, and we restrict the discussion in this section to pointing to interesting examples and useful review materials, which enable a rapid introduction to specific fields.

64.5.1 Road Networks

Road networks are the most immediately obvious network encountered in everyday life. The primal representation of a road network, where vertices are the road intersections and edges are the road segments between them, generally exhibits rather uninteresting structure. Across large areas of a given city, the road network approximates to a two-dimensional lattice, and the range of vertex degrees is limited by geometry: it is unusual for road junctions to connect more than five or six road segments. However, when we range across larger scales, the road hierarchy in most regions introduces shortcuts in the form of highways with limited connections to the lattice of local roads. Spatial small-world networks capture this structure relatively well.

More interesting features of road networks may emerge when the primal representation is converted to the line graph dual or when *named roads* are treated as the units of analysis (i.e., the vertices in the graph) and intersections between named roads are the graph edges. Either of these transformations admits greater variety in the vertex degree distribution, and can enable the topologically most central roads to be identified, which may be of greater interest than more traditional approaches in some cases (see Jiang 2006). Perhaps the most interesting work in this area has been recent efforts to model a variety of urban street networks using rather simple models based on the preferential attachment principle but taking spatial constraints into account (Courtat et al. 2011).

64.5.2 Transport Networks

Transport networks cover a wide range of modes other than road-based. Relative to road networks, the most obvious feature of other modes is their point-to-point or station-connection structure. These features introduce greater potential for departures from planarity, particularly when airline networks and shipping routes

are considered. In analysis of an extensive database of world airline routes, Barrat et al. (2004) demonstrate that this network has many interesting properties, including small-world characteristics, and distinctive scales related to regional and global service areas. This paper and others focusing on airline networks and the various findings in this area are well covered in the spatial networks review paper by Barthélemy (2011, pp. 13–17).

A comprehensive overview of developments in the analysis of all kinds of transport networks is provided by Rodrigue et al. (2009) where coverage extends beyond the more structural forms of analysis discussed in this chapter to cover how transport networks structure the regional and global economy and how they impact urban mobility and related issues of transport policy and planning. The grounding of earlier work in real-world histories provides a striking contrast with recent work in a journal special issue “Evolution of Transportation Network Infrastructure” (see Levinson 2009) where more exploratory analyses of different network growth models are highlighted.

64.5.3 Other Spatially Embedded Networks

It is appropriate given its importance in inspiring much of the recent explosion in work on networks to point to work on the Internet. This is representative of a wide range of work on infrastructure networks of all kinds. It is easy to forget that the Internet relies like other networks on physical plant of various kinds and that considerations such as efficiency of service provision, costs of installation, and vulnerability to disruption are critical concerns for the Internet backbone as they are for other infrastructure such as electricity and water supply. A comprehensive overview of network analysis work on the Internet is provided by Pastor-Satorras and Vespignani (2004). More geographically grounded perspectives that focus attention on the spatial embedding of Internet infrastructure focus on how the structure of the Internet relates to local geographical factors and to other infrastructure networks, often showing that places that are well connected by airlines, roads, and other systems tend also to be well provided with Internet connectivity (Malecki 2002). Once again, the interplay between exploratory analysis of overall structure and more grounded approaches is critical to progress in understanding in this field.

Finally, we briefly consider spatially embedded social networks, perhaps the fundamental building block of all the other networks considered. An excellent overview of how space and social networks may be mutually reinforcing and how these effects can be modeled is provided by Butts and Acton (2011). They strongly argue for the benefits of analysis that attends to both network aspects and spatial aspects. Among the most promising areas for future development in this field are coevolutionary networks (Gross and Blasius 2008) where network structures and the attributes of nodes and edges mutually influence one another over time, and the wide-ranging study of how processes such as disease spread or the diffusion of ideas occur on networks (see Newman 2010, pp. 627–676).

64.6 Conclusions

Many, perhaps most, features of the human world can be considered to be embedded in space and networked to one another at various spatial scales either (more or less) permanently or in constantly changing ways. This chapter has deliberately focused on basic concepts and models that are useful for the analysis of such networks, particularly emphasizing the rapid growth of ideas in the recently emerged “science of networks.” While much of this material has been developed in statistical physics and allied fields, it is apparent that the insights yielded by these approaches build on much earlier work on networks in geography and regional science, extending it and applying fundamental ideas to larger and more dynamic networks than before. Even so, claims that work in these areas heralds a new dawn for the social sciences (see, e.g., Watts 2007) seem overdone. On the contrary, it is probable that the best and most insightful work will continue to demand the application of measures, methods, and models from network science reviewed here, in combination with detailed, well-grounded empirical research on the development and structure of networks in specific contexts in space and time.

References

- Albert R, Jeong H, Barabási AL (1999) Diameter of the world-wide web. *Nature* 401(6749):130–131
- Barat A, Barthélémy M, Pastor-Satorras R, Vespignani A (2004) The architecture of complex weighted networks. *Proc Natl Acad Sci USA* 101(11):3747–3752
- Barthélémy M (2011) Spatial networks. *Phys Rep* 499(1):1–101
- Black W (2003) Transportation: A Geographical Analysis. Guilford Press, New York
- Butts CT, Acton RM (2011) Spatial modeling of social networks. In: Nyerges T, Couclelis H, McMaster R (eds) *The Sage Handbook of GIS and Society Research*. SAGE Publications, Los Angeles, pp 222–250
- Caldarelli G (2007) Scale-free Networks: Complex Webs in Nature and Technology. Oxford University Press, Oxford Finance
- Castells M (1996) *The Rise of the Network Society*. Blackwell, Malden, MA
- Courtat T, Gloaguen C, Douady S (2011) Mathematics and morphogenesis of cities: A geometrical approach. *Phys Rev E* 83:036106
- Erdős P, Rényi A (1960) On the evolution of random graphs. *Publ Math Inst Hung Acad Sci* 5:17–61
- Fischer MM (2004) GIS and network analysis. In: Hensher DA, Button KJ, Haynes KE, Stopher PR (eds) *Handbook of transport geography and spatial systems of handbooks in transport*, vol 5. Elsevier, Kidlington, pp 391–408
- Fortunato S (2010) Community detection in graphs. *Phys Rep* 486(3–5):75–174
- Girvan M, Newman MEJ (2002) Community structure in social and biological networks. *Proc Natl Acad Sci USA* 99(12):7821–7826
- Gross T, Blasius B (2008) Adaptive coevolutionary networks: a review. *J R Soc Interface* 5(20):259–271
- Gross JL, Yellen J (2006) *Graph Theory and its Applications*. Discrete mathematics and its applications. Chapman & Hall/CRC, Boca Raton
- Haggett P, Chorley RJ (1969) *Network analysis in geography*. Edward Arnold, London
- Jiang B (2006) Ranking spaces for predicting human movement in an urban environment. *Int J Geogr Inform Sci* 23(7):823–837

- Jungnickel D (1999) Graphs, networks and algorithms. Springer, Berlin
- Kansky K (1963) Structure of transportation networks: relationships between network geometry and regional characteristics. PhD thesis, Department of Geography, University of Chicago
- Levinson D (2009) Introduction to the special issue on the evolution of transportation network infrastructure. *Netw Spatial Econ* 9:289–290
- Lorrain F, White HC (1971) Structural equivalence of individuals in social networks. *J Math Sociol* 1:49–80
- Malecki EJ (2002) The economic geography of the internet's infrastructure. *Econ Geogr* 78(4):399–424
- Newman M (2010) Networks: an introduction. Oxford University Press, Oxford, UK
- Nordbeck S (1964) Computing distances in road nets. *Pap Reg Sci* 12(1):207–220
- Okabe A, Boots B, Sugihara K, Chiu SN (2000) Spatial tessellations: concepts and applications of Voronoi diagrams, 2nd edn. Wiley, Chichester
- O'Sullivan D, Unwin DJ (2010) Geographic information analysis. Wiley, Hoboken, NJ
- Pastor-Satorras R, Vespignani A (2004) Evolution and structure of the internet: a statistical physics approach. Cambridge University Press, Cambridge, UK
- Rodrigue JP, Comtois C, Slack B (2009) The geography of transport systems, 2nd edn, The geography of transport systems. Routledge, London
- Taylor PJ, Ni P, Derudder B, Hoyler M, Huang J, Witlox F (eds) (2011) Global urban analysis: a survey of cities in globalization. Earthscan, London/Washington, DC
- Wassermann S, Faust K (1994) Social network analysis: methods and applications. Cambridge University Press, Cambridge, UK
- Watts DJ (2007) A twenty-first century science. *Nature* 445(7127):489
- Watts DJ, Strogatz SH (1998) Collective dynamics of 'small-world' networks. *Nature* 393:440–442
- Weiser M (1991) The computer for the twenty-first century. *Sci Am* 265(3):94–104
- Wilson RJ (1996) Introduction to graph theory. Longman, Harlow

Section VIII

Spatial Statistics

spatial

multilevel population space distance
regional distribution interest
mean weight variables
spatially Karthiga
parameter variance
time random
disease GWR
areas applied parameters
health correlation
neighborhood observations
statistical coefficients
area map area map
sampling matrix effects
matrix software estimates
space-time autocorrelation
Gaussian

data regression

matrix effects
estimates autocorrelation
space-time Gaussian

models

Bayesian variable function eigenvectors
values individual geographic

Spatial Data and Statistical Methods: A Chronological Overview

65

Robert Haining

Contents

65.1	Introduction	1277
65.2	Where Did It All Start?	1280
65.2.1	The Statistical Origins	1280
65.2.2	From Statistics into Geography and Regional Science	1281
65.3	Spatial Econometrics	1283
65.4	New Kinds of Geographical Exploration	1286
65.4.1	Exploratory Spatial Data Analysis	1286
65.4.2	The Local Revolution	1287
65.5	Into the Twenty-First Century	1289
65.5.1	Spatial Data Mining	1289
65.5.2	The “New” Geostatistics	1290
65.5.3	Bayesian Hierarchical Modeling	1291
65.6	Conclusions	1292
	References	1293

Abstract

We review some of the special properties of spatial data and the ways in which these have influenced developments in spatial data analysis. We adopt a historical perspective beginning in the early twentieth century before moving to the development of spatial autocorrelation statistics in geography’s Quantitative Revolution. Phases of development after the Quantitative Revolution are divided into emergence of spatial econometrics, the development of exploratory methods for spatial data analysis, and local statistics for handling heterogeneity. We then consider more recent advances in the areas of spatial data mining, the “new” geostatistics, and Bayesian hierarchical statistical modeling of spatial data.

R. Haining

Department of Geography, University of Cambridge, Downing Place, Cambridge, UK
e-mail: rph26@cam.ac.uk

65.1 Introduction

Spatial statistics is used for the *analysis* of spatial data, that is, “the reduction of spatial patterns to a few clear and useful summaries” (Ripley 1981, p. 1), and for *comparing* such summaries “with what might be expected from theories of how the pattern might have originated and developed” (Ripley 1981, p. 1). In order to test theoretical expectations against data collected through observation, statistical models are often necessary. There are, of course, many different types of models that are critical to the progress of science, but statistical models are formalizations of theory in terms of random variables and their associated probability distributions. We might compare several different models to see which one best fits the data; we might take a single model and see how well it fits the data.

There are several different types of spatial data encountered in geography and regional science. *Point pattern* or *point process* data arise where each data value refers to the location of a discrete object the size of which is sufficiently small relative to the study area that it can be treated as a point (e.g., the location of factories in a region). Interest may focus on how the points are distributed within the region (e.g., are the factories in a particular economic sector spatially clustered or random?). If interest focuses on say the distribution of an attribute attached to each point (e.g., are the factories that have been closed in the last 12 months clustered given the distribution of factories in that sector?), we refer instead to a *marked point process*. Some variables that originate as point data are reported as discrete-valued *regional counts* (e.g., the number of residential burglaries recorded by UK Census Output Area (COA)) or their attributes as continuous-valued *regional averages* (e.g., average household disposable income) or *rates* (e.g., the number of burglaries per 100 households) or *ratios* (e.g., area-standardized disease-specific mortality ratios obtained by dividing the observed number of deaths due to a particular disease in each area by the expected number of cases given the area’s population size and its age and sex composition). The reporting areas may be irregular in shape as in the above examples or form a regular grid. When *regional data* refer to small areas (such as COAs or a fine grid), then there may be interest in constructing a density map which is a smoothed representation of the data (e.g., a population density map, a burglary risk map). Some spatial data are *point samples from a continuous surface*, where the point sampling has been performed according to some design (e.g., random, stratified random, systematic). The data may refer to levels of surface soil contamination or ground level atmospheric pollution. There may be interest, for example in *geostatistics*, in constructing a map of the attributes spatial variability or interpolating to points or areas on the map where no data have been collected. Spatial data may also take the form of *objects* that have both location and extent and which may or may not fill the space. Interest may focus on modeling the distribution of the objects such as vegetation or land use patches. As a final example, data may refer to the nodes or vertices of a *network* (e.g., a rail, road, or airline network). On the vertices of the network are recorded origin–destination or (directed) flow data such as numbers of people or tonnage of goods moving between two nodes or line segments on the network in a period of time.

Analysis may be concerned to understand population migration or trade *flow data*, for example, using origin- and destination-specific factors and the distance between the origins and destinations (Fischer and Wang 2011).

In any analysis, it may be necessary to combine data of different types: regular areas, irregular areas, and point data, as illustrated, for example, in Elliott et al. (2009). Different data types raise different problems for statistical analysis and modeling, and for a more formal definition of a number of different forms of spatial data, see Cressie (1991, pp. 8–9)

Obtaining useful summaries of spatial data is complicated because single numbers (such as the mean and standard deviation for summarizing aspects of the distribution of data values) are not sufficient for describing the spatial variation in data values. Maps and graphs, perhaps several of both, are required in order to describe both the distributional and spatial variation in the data. Spatial data may show different patterns of variability on different parts of the map and at different scales.

Statistical analysis and modeling of spatial data introduces other considerations. The classical theory of statistical inference assumes data are obtained by randomly *sampling* from the population (so that the data set can be considered representative of the population), and observations are independent and identically distributed (i.i.d.). For this reason, the underlying probability models of classical statistics are i.i.d. But in the case of spatial data, even if sample observations have been collected by a process of random *sampling*, if sample values are sufficiently close together in geographical space, they will not be independent because the population from which the data have been drawn is said to be *spatially autocorrelated*. Data values are not independent; the structure of that dependence may not be the same everywhere on the map (nonstationarity), and there may also be small pockets of high (or low) data values (*disease clusters*, crime, or unemployment hot spots). The mean value of an attribute may not be the same everywhere on the map. In addition to nonindependence, data values across a map may not be uniform in the sense of coming from some common underlying distribution. They may display what is termed *spatial heterogeneity*. These two often-encountered characteristics of spatial data introduce special considerations when undertaking statistical modeling including the need to construct valid (or “permissible”) models to describe spatial variability which can then be used for inference with spatial data.

When data are collected in the form of counts or rates by area, there may be further issues to consider. If area data values are averages, rates, or ratios and if in addition areas possess large populations, then such summaries may obscure or conceal *within-area heterogeneity* – subpopulations within areas with markedly different averages (or rates or ratios). On the other hand, if the areas possess small populations, while such a framework might preserve more of the underlying spatial variability in the data (and each small area might be homogeneous), individual area estimates will typically have much larger standard errors. As a further consequence of this, if the map is partitioned into areas, some of which have large while others have small populations, then the data may be *heteroscedastic* – that is, each observation has been drawn from a different probability distribution with

a different variance. If area data values refer to an ecological covariate (i.e., expressing a property of an area that is not reducible to measures at the individual level, e.g., social capital, area deprivation, social cohesion), then the values of this variable will depend on the scale of the partition and the configuration of the boundaries (the zoning). This is one example of the “*modifiable areal unit problem*” which states that the results of spatial analyses and modeling are conditional on the particular partition through which we observe spatial variation. These and other challenges that confront the analysis of spatial data have been reviewed at some length elsewhere (see Haining (2009)).

This chapter provides an overview of some of the major developments in spatial statistics with particular relevance to geography and regional science. It is divided into four sections. The first section briefly reviews the statistical origins of what is now called spatial statistics and describes how and why one area of spatial statistics came to be part of geography’s *Quantitative Revolution* in the 1950s, 1960s, and early 1970s. The second section discusses the emergence of *spatial econometrics* and its links (and overlaps) with spatial statistics. What characterizes *spatial econometrics* and in what sense does it stand apart from, indeed distinct from, the field of spatial statistics? The third section considers the emergence of *exploratory spatial data analysis* in the 1980s and follows this with the development of “local statistics” for analyzing spatial *heterogeneity*. The fourth section is an overview of some recent developments in the field: *spatial data mining*, what I shall term the “new” *geostatistics* and Bayesian hierarchical spatial statistical modeling. The purpose of this chapter is to show the reader the development path of spatial statistics and how this can be seen as a response to the distinctive properties and challenges presented by spatial data.

65.2 Where Did It All Start?

65.2.1 The Statistical Origins

The roots of spatial statistics can be traced back to at least the early twentieth century and the involvement of classically trained statisticians in analyzing the data from agricultural uniformity trials carried out at Rothamsted in England. In such analyses, a component of yield variation is due to processes operating at a scale greater than the size of the spatial units used to report the data so that crop yields in adjacent plots tend to be similar. Classically trained statisticians became interested in the problem of how to carry out field trials (which might be testing different management methods and crop varieties in relation to soil type) using experimental designs that would control for such attributes and hence yield stronger inference.

The analysis of agricultural yield data was also motivation for the seminal paper by Whittle (1954) in which he made explicit the link between the problem of analyzing such data and two-dimensional stochastic models. In that paper, he defined the *simultaneous spatial autoregressive (SAR) model* on

a regular square lattice in which if $X(i,j)$ denotes the random variable at location (i,j) of a regular square lattice then

$$X(i,j) = a[X(i-1,j) + X(i+1,j)] + b[X(i,j-1) + X(i,j+1)] + e(i,j) \quad (65.1)$$

where a and b are parameters and $e(i,j)$ is a normally distributed white noise (i.i.d.) process. Whittle's paper notes that unlike time series modeling, spatial process modeling needs to allow for dependence to extend in all directions (not just from past to future as in time series modeling) and that in two dimensions, the dependence structure on the north–south axis might differ from that on the east–west axis. Twenty years later, Besag (1974) in an equally seminal paper presented the theory of conditionally specified spatial models (for continuous- and discrete-valued random variables) and included in his set of models the *conditional spatial autoregressive (CAR) model* for normally distributed random variables. For readers interested in the differences between these models, see, for example, Haining (2003 pp. 297–304).

A slightly earlier strand of development, predating Whittle's work, saw the publication of papers addressing the problem of how to test for what is now referred to as *spatial autocorrelation*. In each case, the null hypothesis of no *spatial autocorrelation* is tested against a nonspecific alternative that the observations are autocorrelated. The tests, *Geary's c test*, based on squared differences between observations in adjacent areas; *Moran's I test*, based on the cross product between observations in adjacent areas; and Krishna Iyer's *join count test* for nominal level data, based on counting the number of adjacent areas in the same class or in different classes, are reviewed in Cliff and Ord (1973). None of these early tests made any allowance for the topological and/or geometrical structure of the areas making up the areal system other than whether pairs of areas were adjacent (shared a common border) or not.

These developments were applicable to the case of area data. *Geostatistics* was developed for spatial data collected as point or block samples from a continuous surface and with a quite different aim in mind. Matheron (1963) developed a comprehensive theory of optimal *interpolation* in geographical space on the basis of sample data. In purely geographical terms, we might think of this as a theory for drawing maps of continuous phenomena on the basis of a scattered sample of observations. Spatial variation in any particular set of data was described by estimating the *semi-variogram* (a squared difference statistic) and then finding a best fit model for this empirical *semi-variogram* from the set of “permissible” models. A rich array of models is available for describing spatial variation. For further discussion of this area of spatial statistics and its antecedents, see Haining et al. (2010) which also includes comparative comments with the literature cited above.

65.2.2 From Statistics into Geography and Regional Science

It was during the *Quantitative Revolution* in the 1950s and 1960s that some aspects of this statistical theory began to filter into geography. Researchers in sociology had

already recognized (in some cases long before geography's *Quantitative Revolution*) that the theory and tools of classical statistics could not be applied uncritically to the analysis of geographical data (see, e.g., Neprash (1934) as well as other papers in the same journal supplement). It was not until the 1960s that the "problem" of *spatial autocorrelation* began to be examined carefully by geographers and a key deficiency of the earlier tests devised by Moran, Geary, and Krishna Iyer, their topological invariance, confronted.

Geography's *Quantitative Revolution* was not merely a methodological revolution – the aligning of geography's methods with those used by the quantitative sciences – it was also a revolution in terms of how the subject matter of geography should be addressed. It was a revolution in the sense that geographers became interested in the development of theory. But it would be theory that would only survive so long as it withstood rigorous attempts to refute it through empirically grounded research. Models represented the translation of theory into a form that would enable empirical testing to be performed, and although not the only form of model building that entered the geographical literature, statistical modeling was a key element in this agenda. So herein lay the nub of the problem. Because of the importance of statistical modeling to theoretical geography, these issues could not be set to one side. And there were of course precedents to believing that none of these problems were insuperable, least of all coping with the effects of spatial dependence. Time series analysis had undergone a transformation in the first half of the twentieth century and now underpinned the practice of econometrics which in turn supported the testing of economic theory. Geographical statistics was in need of a similar transformation.

It was the work of Cliff and Ord that reported important breakthroughs in constructing statistics for testing for *spatial autocorrelation* on the sorts of irregular areal frameworks social scientists most frequently worked with. In Cliff and Ord (1973), they developed the inference theory for modified versions of *Geary's c* and *Moran's I* statistics introducing a "weighting" term into the formulation of the statistic that allowed adjacency to be specified much more generally than had hitherto been possible. The reader interested in this aspect of geography's history should refer to the special issue of the journal *Geographical Analysis* (2009(4)).

Two areas of geography benefitted most directly from this innovative work. The first was in the application of the regression model. *Regression models* enable data analysts to empirically test the relationship between a dependent variable and a set of explanatory or independent variables. This statistical model has in the past and indeed continues to play a very important role in developing and testing theory in many areas of quantitative science. In the case of classical least squares *regression modeling*, population inference (hypothesis testing, parameter estimation) is based on the assumption that model errors are i.i.d., and failure to satisfy this assumption results in underestimation of type I errors in hypothesis testing. Regression residuals are estimates of model errors, and Cliff and Ord (1973) provide an inference theory for testing for nonindependence of model errors using the least squares residuals.

The second area to benefit was the testing of specific types of spatial theory. Cliff and Ord (1973) considered the area of spatial *diffusion modeling* where different theoretical processes were simulated and then compared with observed outcomes. But, among many other complications associated with this approach, evaluating the correspondence between simulated output and empirical observation must compare not only the frequency distribution of numbers of adopters by area but also the spatial arrangement of the counts.

Another example was in the area of economic geography and the analysis of those economic processes that by definition are embedded into geographical space (e.g., urban and regional development, location theory, land use change, regional and international trade, spatial price competition). During the *Quantitative Revolution*, those economic geographers who went in search of stronger theoretical perspectives began to take a close interest in the work of location theorists. They also began to engage with the newly emerging field of regional science. Early books on the tools and methods of regional science paid little attention to statistical modeling, but in the 1970s, the field of *spatial econometrics* began to emerge becoming to regional scientists what econometrics had become to economists. The purpose behind its development was to provide the statistical tool kit to enable regional scientists to test spatial economic theory. We now turn to discuss this development.

65.3 Spatial Econometrics

Anselin (1988), and again most recently in his extended review of the field in 2010, credits Jean Paelinck with first use of the term *spatial econometrics* in an address to the Dutch Statistical Association in 1974. The term was used to “designate a growing body of the regional science literature that dealt primarily with estimation and testing problems encountered in the implementation of multiregional econometric models” (Anselin 1988, p. 7). Another significant date in the development of the field is 1979 when Paelinck and Klaassen’s (1979) “*Spatial Econometrics*” was published. Anselin (2010) chooses that year as “the historical starting point for *spatial econometrics*” (p. 3).

Anselin (2010), as others had done before him, argues that *spatial econometrics* and spatial statistics should be seen as distinct. The distinction is defined by the types of problems that are tackled. Whereas spatial statistics is fundamentally data driven, *spatial econometrics* (like econometrics), is fundamentally theory driven. *Spatial econometrics* has been developed explicitly to fit *spatial regression models* to test spatial economic theory – in this sense moving away from Paelinck’s original definition of the field.

Providing spatial econometricians do not cut themselves off from the rich vein of statistical theory and models generated by spatial statisticians, then there may be advantage to be gained from distinguishing between *spatial econometrics* and spatial statistics. But the justification for the distinction is not entirely convincing.

In contrast to earlier days in geography's *Quantitative Revolution*, statisticians today see many opportunities for fruitful interaction on broad classes of spatial problems and would not accept the view that their model building is purely data driven, by implication "atheoretical." As Cressie and Wikle (2011, p. 14) have recently observed, "...Statistics has become more a Science than a branch of Mathematics..."

Spatial econometrics today is principally concerned with how to specify, fit, and then carry out diagnostic checks on *regression models* when working with locationally (or spatially) referenced data. The data can be cross-sectional (purely spatial) or *spatiotemporal* with measurements on one or several variables. Underlying these models are usually theories about how distance (to a particular location such as a city center), spatial configuration (the spatial distribution of objects within a space, such as whether areas of poverty are scattered or ghettoized within an urban area), or spatial gradients (between neighboring areas in terms of socio-economic characteristics) help to explain variation in a dependent variable. What is observed (e.g., area crime rates, regional economic performance) is not necessarily an outcome purely of circumstances *within the places themselves* because what is observed, and the variation we want to explain, may be the outcome of processes that operate across geographical space (e.g., different forms of interaction).

Methodologically *spatial econometrics* focuses on two properties commonly encountered when handling geographical data: spatial dependence (autocorrelation) and spatial *heterogeneity*. We shall consider approaches to the handling of spatial *heterogeneity* in the next section and focus here on just the handling of spatial dependence in *spatial econometrics*.

Typically, spatial dependence is handled by specifying lagged variables in the *regression model*. In the case of lagging the dependent variable, a model might be specified of the form

$$Y(i) = b_0 + b_1 X_1(i) + \dots + b_k X_k(i) + \rho \sum_{j \in N(i)} w(i,j) Y(j) + e(i) \quad (65.2)$$

$$\sum_{j=1}^n w(i,j) = 1; \quad i = 1, \dots, n$$

where the $\{e(i)\}$ are i.i.d $N(0, \sigma^2)$, b_0 is the intercept coefficient, and b_1, \dots, b_k are the regression coefficients on the independent variables X_1 to X_k . The parameter ρ is the spatial interaction parameter for the weighted average of the dependent variable ($Y(i)$). For any given site i , this weighted average takes in the values at sites neighboring i but excluding site i ($N(i)$). Thus, $w(i,j) > 0$ if $j \in N(i)$ and $w(i,i) = 0$ for all i . Models specified in this way and in which the influence of neighboring sites is usually stronger the closer they are to i ($w(i,j) > w(i,k)$ if j is closer to i than k is to i) have a long history in the statistical modeling of certain types of economic interaction processes including price competition effects. Clearly, other forms of weighting could be constructed to reflect the structure of economic interactions across space (Haining 1990).

Lagging may also be specified on one or more of the independent variables as in, for example, the model

$$\begin{aligned} Y(i) &= b_0 + b_1 X_1(i) + \dots + b_k X_k(i) + b_{r,\text{lag}} \sum_j c(i,j) X_r(j) + e(i) \\ \sum_{j=1}^n c(i,j) &= 1; \quad i = 1, \dots, n \end{aligned} \quad (65.3)$$

where in this case $Y(i)$ is modeled to have an association with the independent variable X_r as a function not only of X_r 's value at i but also its value at neighboring locations. (For this reason, we use the notation $c(i,j)$ (rather than $w(i,j)$) to distinguish the spatial averaging in Eq. (65.3) from that in Eq. (65.4).) This type of model is sometimes encountered in house price modeling where characteristics of the neighborhood where the house is located and nearby neighborhoods may impact on price. This type of spatial averaging or smoothing may also be encountered in environmental epidemiology where the air pollution level in neighboring areas is treated as a risk factor because people move about in their day-to-day lives and are thus exposed to levels of air pollution in areas other than where they reside.

In the absence of well-defined explanatory variables to include in the model, the spatial lagging may be applied to the errors

$$\begin{aligned} Y(i) &= b_0 + b_1 X_1(i) + \dots + b_k X_k(i) + u(i) \\ u(i) &= \theta \sum_{j \in N(i)} w(i,j) u(j) + e(i) \\ \sum_{j=1}^n w(i,j) &= 1; \quad i = 1, \dots, n \end{aligned} \quad (65.4)$$

where the terms in Eq. (65.4) are as defined above and θ is now the spatial interaction parameter associated with the errors. Forms of this model and Eq. (65.2) have been used in the modeling of origin–destination *flows* (Fischer and Wang 2011, pp. 64–67).

As Anselin (2010) points out, the methodology associated with the fitting of this class of models has continued to evolve. He reviews in some detail the notable strides that have been made both in the rigor with which these and other models can be fitted and in the availability of software to implement the fitting. One problematic aspect in this evolutionary development is the specification of the *weights matrix* $\{w(i,j)\}$. Adjacency is the default option for many analysts in specifying the *weights matrix* with two areas being defined as *neighbors* if they share a common border. Most software makes this an easy option to implement. But adjacency may not always be appropriate depending on the model to be fitted and whether, for example, there is a need to capture other forms of spatial relationship including

hierarchical dependency structures and complex patterns of spatial competition (Haining 1990). Another approach is to define the elements of the *weights matrix* based on the similarity of the areas in terms of one or more covariates when borrowing data spatially to strengthen small area inference, for example, social and other interpersonal networks may be used to underpin spatial relationships based on the presence or absence of social relationships. Lu et al. (2007) define, as part of a Bayesian *hierarchical model*, an intrinsic conditional autoregressive model where the weights, $w(i,j)$, are Bernoulli distributed with parameter $p(i,j)$ and where $\text{logit}(p(i,j))$ is a linear function of a set of covariates ($z(i,j)$) based on known features of the pair of areas i and j .

The links between methodological advance and the evolution of spatial economic theory are only touched upon in Anselin (2010) – in that sense, his review is concerned with theoretical *spatial econometrics* (statistical methods) rather than applied *spatial econometrics* (economic models). Over time, applied *spatial econometrics* has tended to become synonymous with *regression modeling* applied to spatial data where *spatial autocorrelation* and *spatial heterogeneity* in particular are present and need to be accommodated. Its treatment of spatial effects reflects the growing “legitimization of space and geography” (Anselin 2010, p. 8) in the quantitative social sciences more generally. But the subfield perhaps needs to be more than that if it is to justify its separate identity from spatial statistics and fully justify its “econometric” label. A close link with mainstream economic theory would seem essential in order to provide economic legitimacy to models (systems of equations) within which geography and spatial relationships have been, in economic terms, rigorously embedded (Fingleton (2000)).

65.4 New Kinds of Geographical Exploration

65.4.1 Exploratory Spatial Data Analysis

Exploratory data analysis (EDA) is a collection of techniques for summarizing data properties, detecting patterns in data, identifying unusual or interesting features in data, detecting errors, distinguishing accidental from important features in a data set, and formulating hypotheses from data. *EDA* might also be used in later phases of analysis, for example, in assessing model fit. Techniques are typically visual (charts, graphs, and figures) and/or numerical (resistant statistics, i.e., statistics not greatly affected by a small number of extreme values). *Exploratory spatial data analysis* (ESDA) extends the definition of *EDA* to spatial data, extending the set of visual tools to include the map and the set of numerical tools to include, for example, spatial *cluster detection* statistics (Haining 1990, 2003).

GIS or *GIS-like* software, for example, GeoDa, have provided excellent platforms for these tools (see www.geodacenter.asu.edu). Advances in computer technology have had a particularly big impact on *ESDA* with the development of new visualization techniques such as *brushing* (highlighting cases in one graph such as a segment of a boxplot and seeing them highlighted in another graph or on a map),

dynamic brushing (brushing using a moving window), and various forms of *dynamic interactivity* (allowing the user to modify the graphics themselves to better explore data properties, e.g., rescaling and rotating three-dimensional plots).

A significant challenge in undertaking these forms of analysis with area data (where data values refer to areal aggregates such as census tracts) is the problem of comparability especially when dealing with small numbers of events (e.g., numbers of cases of a disease by small area). In many areas of *scientific visualization*, different data values are directly comparable (e.g., the results of experiments, data values taken from time windows of the same length). But area data across a region often refer to polygons of different physical sizes and with different baseline populations. This raises two distinct problems for *ESDA*. Comparing rates and ratios across such a map is potentially misleading. A rate computed for an area with a small denominator population has a larger error variance than a rate computed for an area with a large denominator population. This may necessitate using different sized symbols or other visual devices to distinguish between high- and low-precision data values. Extreme rates (and ratios, such as standardized ratios where an observed count is divided by an expected count) are most often found when the denominator population is small, but statistically significant rates (relative to some baseline) are most often found when the denominator population is large (see, e.g., Haining 2003, pp. 194–199). Sometimes, the areas with the largest populations are physically the smallest (e.g., the census tracts in urban as compared to rural areas) and may be hard to see depending on the scale of the map. One solution to this is the cartogram, where, for example, each area is physically transformed so that its size is proportional to its population. As computer technology has advanced, it has become possible to develop many different forms of cartogram, some that more closely reflect the area as the viewer is used to seeing it which may help him or her to better navigate and hence read the map. For numerous examples of *cartograms* see, for example, www.worldmapper.org and www.sasi.group.shef.ac.uk/maps.

65.4.2 The Local Revolution

We noted above that spatial *heterogeneity* is a property often present particularly when analyzing spatial data over a large geographic area. *Heterogeneity* may be illustrated in the following terms. Assume we are dealing with the counts associated with the number of new cases of a disease in each of n areas during an interval of time. Suppose the generating process for these counts is dependent on an underlying set of risk parameters $(\lambda_1, \dots, \lambda_n)$. If all the λ_i are identical, then the risk surface is said to be homogeneous. If for at least some areas ($i \neq k$), $\lambda_i \neq \lambda_k$, then the risk surface is said to be spatially heterogeneous.

Clusters of cases might arise from a process in which events occur independently of each other but where there is spatial variation in the levels of the risk factors across the map. So, an unusually large number of observed cases in area k may be the product of a high value of λ_k which may be due, in turn, to high levels of the relevant

risk factors in area k that determine the value of λ_k . If adjustment is made for these factors, then the existence of the cluster may be accounted for. Clusters of cases might also be due to a contagion process where although the underlying risk map may be uniform, when one case occurs, it triggers others giving rise to a spatially clustered pattern. Examples of this type of process include repeat offending in the same neighborhood for cases of burglary and the occurrence of cases of an infectious disease. This is referred to as *global heterogeneity*, extra variation in the data that can be analyzed using a global model. This *heterogeneity* may be spatially uncorrelated, but it may be spatially correlated if, for example, the spatial scale of the process exceeds the size of the observational units used to collect data.

However, there is often interest in identifying the specific locations of local clusters (hot spots) of cases in an area referred to as *local heterogeneity*. Kulldorff's scan test, a likelihood-based test statistic, has been widely adopted to test for the presence of spatial clusters in point as well as area data (Kulldorff 1997). This test uses moving windows (circles) of varying size. Each of these many circles represents a possible cluster. The test measures the unusualness of each potential cluster using a local likelihood ratio statistic which compares a null hypothesis that cases occur in the population at risk with equal probability whether individuals are inside or outside the circle against an alternative hypothesis that cases inside the circle have a higher probability of occurrence than those outside the circle. The circle with the highest local likelihood ratio statistic is considered the most likely cluster. The question posed is "how unusual is this most unusual collection of events?" (Waller 2009, p. 312). By using Monte Carlo hypothesis testing, the scan test is able to answer this question avoiding the *multiple testing problem*. In addition to the scan test that looks for clusters wherever they might be on the map, another class of techniques tests for whether there is an unusually large number of cases around a specific location such as a point source of pollution or other source of possible contamination. This is referred to as a *focused test* (see, e.g., Haining 2003, pp. 263–5).

That *heterogeneity* may be present in the relationship between a dependent variable and the set of independent variables that explain its spatial variation underlies another important set of local spatial statistical techniques. Such a *regression model* might take the form

$$Y(i) = b_0 + b_1(i)X_1(i) + \dots + b_k(i)X_k(i) + e(i) \quad i = 1, \dots, n \quad (65.5)$$

where the terms are as defined for Eq. (65.2), but now the regression coefficients depend on i so that parameter values differ for each observation. Additional modeling assumptions have to be introduced in order to fit a model of this type as otherwise there is insufficient data to estimate the parameters.

Consider the hedonic *regression modeling* of house prices in an area large enough to encompass different climatic regimes so that house buyers attach different values to housing attributes depending on location with respect to these different regimes. If the geographic area can be partitioned into different areas, then a *spatial regimes* model may be used in which model parameters are allowed to differ

from one area to another (but not within areas) using dummy variables in the regression that distinguish between the different areas (Anselin 1988). Spatial *heterogeneity* however might show a form of variation where the parameters vary continuously across the study area, rather than discretely, preventing any prior partitioning. In this case, other methods might be implemented, and the interested reader is referred to Jones and Casetti (1992) for the *expansion method* and to Fotheringham et al. (2000) for the method of *geographically weighted regression (GWR)*. In the case of the *expansion method*, the parameters are expressed as a function of a finite number of other variables called expansion variables (z). To take a simple example of the *spatial expansion method* where parameter variation is treated as a function of spatial location, then we might assume

$$b_1(i) = \varphi_0 + \varphi_1 z_1(i) + \varphi_2 z_2(i) \quad i = 1, \dots, n \quad (65.6)$$

where φ_1 and φ_2 are parameters and the variables z_1 and z_2 are the coordinates defining the centroid of each area i . In this case, the spatial expansion is a linear or first-order trend surface which is the additional modeling assumption about how the regression parameters vary spatially. Higher-order trend surfaces could be used or indeed other types of variables. By contrast, *GWR* is based on obtaining local estimates of each parameter where a separate model is fitted to each area. For any $b_j(i)$, for example, data at i are used as well as data from areas close to i but giving most weight to those data values nearest to i . Many possible weighting functions (spatial kernels) can be specified. The additional modeling assumption is that the data in nearby areas to any i carry information about the value of the parameter in i (a form of *spatial autocorrelation*). For a comparative overview of these and other methods for allowing local variation in *regression model* parameters, including *spatially varying coefficients models*, see Lloyd (2011, pp. 109–143).

Heterogeneity may be associated with other properties of the attribute such as its spatial dependency structure. The spatial dependency structure might be different on different parts of the map. Again, depending on the nature of the spatial correlation, there may be a single, global, model of spatial variation that can accommodate the apparent *heterogeneity*. However, there may be circumstances where a global model of spatial variation will not provide a useful model for the data, for example, where there is either theoretical or empirical evidence (or both) that data from particular parts of a map reflect the outcome of special and distinctive local processes. In *geostatistics*, for example, different *variograms* may be needed for different map segments in order to implement kriging.

65.5 Into the Twenty-First Century

In this section, we reflect on some of the areas of spatial statistics which are shaping and will probably continue to form a significant part of the research agenda in spatial statistics in the coming years. We look at the following areas: *spatial data mining*, the “new” *geostatistics*, and *Bayesian spatial hierarchical modeling*.

65.5.1 Spatial Data Mining

Spatial data mining is the process of discovering interesting but potentially useful patterns in spatial databases. It therefore shares at least some of the objectives of ESDA described in an earlier section. But *spatial data mining* is concerned with the development of automated methods that can be applied to large (and very large) spatial databases. Extracting patterns from large databases underpins decision making in many organizations including those concerned with public health, crime and disorder, land use and transportation, and environmental management.

In common with the relationship of ESDA to EDA, *spatial data mining* when compared to other forms of (nonspatial) *data mining* has the special challenge of recognizing spatial relationships and spatial *neighbors* and taking into account the special properties of spatial data. The location and spatial extension of objects need to be embedded into algorithms. “Neighbor relations” need to be examined for many objects within the same analysis and the term “neighbor” interpreted in many different ways for a thorough interrogation. Moreover, given the size of databases and hence the time taken to process data, it has to be possible to achieve efficient implementation for the purposes of, among others, detecting spatial clusters, spatial outliers and co-location, and relationship patterns among different classes of point, line, and polygon objects such as the distribution of an animal species and wildlife habitats. This is one aspect of the “process of stimulus and convergence” between Geographic Information Systems (*GIS*) and spatial data analysis which began in the 1960s and discussed by Goodchild and Haining (2004): “it is more difficult to analyse the vast amounts of (spatial) data available . . . , and to test new theories and hypotheses without computational infrastructure; and the existence of such infrastructure opens possibilities for entirely new kinds of theories and models, and new kinds of data” (p. 382). *GIS* has an important role to play in providing the necessary computational infrastructure for *spatial data mining*. For further discussion as well as numerous examples of *spatial data mining*, see Miller and Han (2009).

65.5.2 The “New” Geostatistics

Traditionally, *geostatistics* has been viewed as a tool to enable physical and environmental scientists to analyze sample data obtained from a continuous surface. But more recently, the methods of *geostatistics* have been adapted to predict and map *regional data* in the form of small area counts. Oliver et al. (1998) use *binomial co-kriging* to analyze the risk of childhood cancer in the English West Midlands. Population size variation across the areal units is taken into account with pairs of areas with larger populations (and hence more reliable rates) given more weight in the estimation of the *variogram*. If the population at risk is large and the probability of having the disease is small so that the *small number problem* arises, *Poisson kriging* can be used.

Geostatistical change of support methods have been used to create maps that help to reduce the visual bias that can arise when mapping data where the subareas

vary in physical size. Areas that are physically large can visually dominate a map. The methodology involves deconvolution of the *variogram* obtained from areal data in order to construct a point support *variogram*. Area-to-point kriging is used to provide point support predictions. Population size variation is allowed in estimating the deconvoluted *variogram*. See Haining et al. (2010) for many references on the methodology including area to area kriging for irregularly shaped areas which can be used to tackle other change of support problems.

Geostatistics is also being used to model spatial variation in a dependent variable in terms of a set of independent variables where the data refer to irregular areas. Kerry et al. (2010) use the spatial components from area to area factorial *Poisson kriging* to identify the most important spatial scales at which crime rates vary and to identify which explanatory variables are statistically significant at those different scales. This represents another important extension of *geostatistical* theory, one that offers insights into the scale-dependent nature of relationships.

65.5.3 Bayesian Hierarchical Modeling

We conclude this section with comments on some new approaches to modeling spatial data. In the last 10–15 years, *Bayesian models* have emerged as important tools in geography and regional science research, made possible by important breakthroughs in computational methods and the availability of inexpensive high-speed computers and of software for fitting spatial models (e.g., WinBUGS and facilities in R and MATLAB). In earlier years, spatial modeling was overwhelmingly *frequentist* or likelihood based: data values, \mathbf{x} , are assumed a random sample from \mathbf{X} , a random variable with a specified probability distribution depending on a set of fixed parameters, $\boldsymbol{\phi}$. The likelihood function for $\boldsymbol{\phi}$ given the data \mathbf{x} is then defined, $L(\boldsymbol{\phi}|\mathbf{x})$, and parameter estimates are based on maximizing the likelihood function. Hypothesis testing is based on likelihood ratios for different values of $\boldsymbol{\phi}$. Inference is based on repeated *sampling*.

With *Bayesian* inference, however, the parameters are also random variables with their own distribution. This means that in addition to specifying the distribution of \mathbf{X} for the observed data \mathbf{x} , it is necessary to also specify the distribution of $\boldsymbol{\phi}$, called the prior distribution, which depends on a further set of parameters. These parameters in turn can be modeled by prior distributions (hyper-priors). The combination of these conditional distributions produces the posterior distribution, and by *sampling* the posterior distribution, inference summaries can be obtained such as the posterior mean, credible intervals (the *Bayesian* version of the *frequentist's* confidence interval), and probabilities of interest (such as the probability of a risk parameter exceeding a critical threshold). In *Bayesian* analysis, instead of handling spatial dependency effects in the data model for \mathbf{X} , which complicates the likelihood and often makes model fitting by maximum likelihood difficult (for an early discussion of this, see Whittle 1954), these effects can be handled in the prior distribution instead and fitted using the software referred to above. There are now many examples of this type of modeling (see, e.g., Le Sage 2000; Lu et al. 2007).

Specifying probability models in terms of a sequence of linked conditional models offers a means of modeling complex systems in ways that quantify the inherent uncertainties within scientific research relating to the data (level 1), the specification of the model (level 2), and model parameters (level 3). “Hierarchical statistical modelling represents a way to express uncertainties through well defined levels of conditional probabilities” (Cressie and Wikle 2011, p. 15). Cressie et al. (2009) provide a discussion and application of *hierarchical models* in *ecological analysis*.

Spatial effects are typically handled through a spatially structured random effects term as the following example illustrates. Suppose the researcher is modeling small area disease counts where $x(i)$ is the number of cases in area i . The data model (level 1) specifies $x(i)$ as the realization of a Poisson random variable ($X(i)$) with intensity parameter $\lambda(i) = E(i)\theta(i)$ where $E(i)$ is the number of cases expected in area i given its population composition and $\theta(i)$ is the area-specific relative risk in area i . This level of the *hierarchical model* expresses the uncertainty in the data given the model specification including its parameters. At level 2, we define the model that reflects our understanding of what determines area level relative risk. For example, we may set

$$\text{Log}[\lambda(i)] = \text{Log}[E(i)] + b_0 + b_1 Z_1(i) + \dots + b_k Z_k(i) + u(i) + s(i) \quad i = 1, \dots, n \quad (65.7)$$

where Z_1, \dots, Z_k define a set of k area-specific covariates with parameters b_1 to b_k that explain variation in relative risk and $\{u(i)\}$ and $\{s(i)\}$ are random effects. The $\{u(i)\}$ are i.i.d. normal random effects, and the $\{s(i)\}$ are given an *intrinsic conditional spatial autoregressive* (ICAR) specification (Haining 2003). These two terms model the scientific uncertainty in the model specification (e.g., competing theoretical understandings of the determinants of relative risk, our understanding of exposure to risk factors) as well as the effects of *overdispersion* and *spatial autocorrelation* in the spatial variation in relative risk and hence in the spatial distribution of the counts. At level 3, for a fully *Bayesian* analysis, the parameters at level 2 are treated as random variables and given probability distributions. As noted in an earlier section of this chapter, this could be extended to include the weights that define which areas are treated as *neighbors* in the ICAR specification. Choices about probability distributions could be informed by scientific understanding, but they might also be a way of allowing for uncertainty in our knowledge (Cressie et al. 2009). For an extension of these models to the multivariate case including multivariate spatial effects, see, for example, Gelfand and Vounatsou (2003).

65.6 Conclusions

One of the earliest items on the agenda of the USA’s National Center for Geographic Information and Analysis (NCGIA) was spatial *data quality* emphasizing its fundamental importance to the development of good science. Understanding data uncertainty, arising from all the stages by which a complex geographical

reality is translated into spatial data, remains at the heart of good spatial science. In the light of the preceding comments, it should also link closely with modeling. At about the same time, attention was also being drawn to the importance of software development for spatial data analysis. In addition to progress in these two areas, the field of spatial data analysis has grown in many other ways. But static, cross-sectional in time, spatial data analysis is restricted to analyzing and modeling the “here and now” of some wider process. A series of spatial analyses over time can shed light on change but in other respects remains limited. The understanding that has been gained by the progress made in spatial statistics forms an essential element in the emergence of *spatiotemporal* data analysis. With the huge growth in space-time data sets and the potential they offer to advance scientific understanding, this represents one of the key areas for future growth.

References

- Anselin L (1988) Spatial econometrics: methods and models. Kluwer, Dordrecht
- Anselin L (2010) Thirty years of spatial econometrics. *Pap Reg Sci* 89:3–25
- Besag J (1974) Spatial interaction and the statistical analysis of lattice systems. *J R Stat Soc, B* 36:192–225
- Cliff AD, Ord JK (1973) Spatial Autocorrelation. Pion, London
- Cressie N (1991) Statistics for spatial data. Wiley, New York
- Cressie N, Wikle C (2011) Statistics for spatio-temporal data. Wiley, New York
- Cressie N, Calder CA, Clark TS, Ver Hoef JM, Wikle CK (2009) Accounting for uncertainty in ecological analysis: the strengths and limitations of hierarchical modelling. *Ecol Appl* 19:553–570
- Elliott P, Richardson S, Abellán JJ, Thomson A, de Hoogh C, Jarup L, Briggs DJ (2009) Geographic density of landfill sites and risk of congenital abnormalities in England. *Occup Environ Med* 66:81–89
- Fingleton B (2000) Spatial econometrics, economic geography, dynamics and equilibrium: a ‘third way’? *Environ Plan A* 32:1481–1498
- Fischer M, Wang J (2011) Spatial data analysis: models, methods and techniques. Springer, Heidelberg
- Fotheringham S, Brunsdon C, Charlton M (2000) Quantitative geography: perspectives on spatial data analysis. SAGE, London
- Gelfand A, Vounatsou P (2003) Proper multivariate conditional autoregressive models for spatial data. *Biostatistics* 4:11–25
- Goodchild MG, Haining RP (2004) GIS and spatial data analysis: converging perspectives. *Pap Reg Sci* 83:363–385
- Haining RP (1990) Spatial data analysis in the social and environmental sciences. Cambridge University Press, Cambridge
- Haining RP (2003) Spatial data analysis: theory and practice. Cambridge University Press, Cambridge
- Haining RP (2009) The special nature of spatial data. In: Fotheringham AS, Rogerson PA (eds) The SAGE handbook of spatial analysis. SAGE, Los Angeles, pp 5–24
- Haining RP, Kerry R, Oliver M (2010) Geography, spatial data analysis and Geostatistics: an overview. *Geogr Anal* 42:7–31
- Jones JP III, Casetti E (1992) Applications of the expansion method. Routledge, London
- Kerry R, Goovaerts P, Haining RP, Ceccato V (2010) Applying geostatistical analysis to crime data: car-related thefts in the Baltic States. *Geogr Anal* 42:53–77

- Kulldorff M (1997) A spatial scan statistic. *Commun stat: theory methods* 26:1481–1496
- Le Sage J (2000) Bayesian estimation of limited dependent variable spatial autoregressive models. *Geogr Anal* 32:19–35
- Lloyd CD (2011) Local models for spatial analysis. CRC Press, Boca Raton
- Lu H, Reilly CS, Banerjee S, Carlin B (2007) Bayesian areal wombling via adjacency modelling. *Environ Ecol Stat* 14:433–452
- Matheron G (1963) Principles of geostatistics. *Econ Geol* 58:1246–1266
- Miller H, Han J (2009) Geographic data mining and knowledge discovery. CRC Press, Boca Raton
- Neprash JA (1934) Some problems in the correlation of spatially distributed variables. *J Am Stat Assoc* 29(suppl):167–168
- Oliver MA, Webster R, Lajaunie C, Mann JR, Muir KR, Parkes SE, Cameron AH, Stevens MCG (1998) Binomial cokriging for estimating and mapping the risk of childhood cancer. *Math Med Biol* 15:279–297
- Paelinck J, Klaassen L (1979) Spatial econometrics. Saxon House, Farnborough
- Ripley BD (1981) Spatial statistics. Wiley, New York
- Waller LA (2009) Detection of clustering in spatial data. In: Fotheringham AS, Rogerson PA (eds) The SAGE handbook of spatial analysis. SAGE, Los Angeles, pp 299–320
- Whittle P (1954) On stationary processes in the plane. *Biometrika* 41:434–449

Jürgen Symanzik

Contents

66.1	Introduction	1295
66.2	Types of Spatial Data	1297
66.3	Basic Visualization and Exploration Techniques via Maps	1297
66.3.1	Choropleth Maps	1298
66.3.2	Linked Micromap Plots	1299
66.3.3	Conditioned Choropleth Maps	1301
66.4	ESDA via Linking and Brushing	1302
66.5	Local Indicators of Spatial Association (LISA)	1304
66.6	Software for ESDA	1306
66.6.1	ESDA and GIS	1306
66.6.2	Stand-Alone Software for ESDA	1306
66.7	Conclusions	1307
66.8	Cross-References	1308
	References	1308

Abstract

In this chapter, we discuss key concepts for exploratory spatial data analysis (ESDA). We start with its close relationship to exploratory data analysis (EDA) and introduce different types of spatial data. Then, we discuss how to explore spatial data via different types of maps and via linking and brushing. A key technique for ESDA is local indicators of spatial association (LISA). ESDA needs to be supported by software. We discuss two main lines of software developments: GIS-based solutions and stand-alone solutions.

J. Symanzik

Department of Mathematics and Statistics, Utah State University, Logan, UT, USA
e-mail: juergen.symanzik@usu.edu; symanzik@math.usu.edu

66.1 Introduction

In his groundbreaking book from 1977 on exploratory data analysis (EDA), Tukey (1977) made several statements that are still relevant today, more than 35 years after the publication of this book:

- “The greatest value of a picture is when it *forces* us to notice what we never expected to see.” (p. vi)
- “Today, exploratory and confirmatory can — and should — proceed side by side.” (p. vii)
- “Exploratory data analysis is detective work — numerical detective work — or counting detective work — or graphical detective work.” (p. 1)
- “Unless exploratory data analysis uncovers indications, usually quantitative ones, there is likely to be nothing for confirmatory data analysis to consider.” (p. 3)
- “Exploratory data analysis can never be the whole story, but nothing else can serve as the foundation stone — as the first step.” (p. 3)

Tukey’s expectations (and limitations) on EDA can easily be extended to exploratory spatial data analysis (ESDA), that is, the exploratory analysis of data with a spatial (geographic) component. As early as 1981, Ripley (1981) followed the distinction between exploratory and confirmatory data analyses in the preface of his book on spatial statistics: “The techniques presented are designed for both of John Tukey’s divisions of exploratory and confirmatory data analysis” (p. vi).

The exact definition of ESDA slightly differs from source to source, but all agree that ESDA is an obvious extension of EDA. Commonly found topics that are covered by ESDA include the visualization and exploration of data in a spatial (geographic) framework. ESDA utilizes many methods, tools, and software components from the field of interactive and dynamic statistical graphics, such as brushing and linked views/linked windows. Typically, one or more map views are linked with one or more statistical displays of the data. Modifications to one of the views will result in modifications of all linked views. Questions of interest that ESDA can answer may be whether a cluster of points that can be seen in a scatterplot is related to nearby spatial locations or whether a particular geographic region (say, the coastal region of a country) exhibits different characteristics than the mountainous region that can be seen in a linked statistical view. Moreover, ESDA can help to create new hypotheses about the underlying spatial data that can later be investigated in more detail in a follow-up study. Also, ESDA methods should be applied before any advanced modeling and testing of statistical hypotheses. Anscombe (1973) has provided some striking examples which could happen when a linear regression line is blindly fitted to some unsuitable data set. The same is the case if some methods from spatial statistics are blindly applied to some spatial data set when no prior exploration took place. We should keep in mind that spatial data often are large and diverse data sets, rather than homogeneous data sets. A large number of different methods usually could, and should, be used, including, simple numerical summary statistics. Coming back to Tukey, the goal or expected outcome of the exploration usually is unknown in advance. Moreover, it should be noted that ESDA is more than just an extension of EDA as additional techniques

and methods are needed that incorporate the specific spatial structure of the data. A frequent goal of ESDA is the exploration of spatial autocorrelation. We speak of positive spatial autocorrelation when nearby observations on average are more similar than what a random assignment would yield and of negative spatial autocorrelation when nearby observations on average are more distinct than what a random assignment would yield.

In the next section, we will discuss the main types of spatial data and following that, consider basic visualization and exploration techniques via maps. We then discuss two of the key concepts of ESDA: exploration via linking and brushing and local indicators of spatial association (LISA). A section on software for ESDA follows. We then finish with a brief conclusion and outlook on possible future work.

66.2 Types of Spatial Data

There exist four main types of spatial data:

- a. In spatial point patterns, the location of an event is of interest itself. Point patterns can be the locations where a patient died from a particular disease or where some specific animal species has been observed. A question of interest might be to explore the spatial patterns of the deaths, for example, at which locations deaths have been due to disease A and at which locations deaths have been due to disease B.
- b. Lattice data, sometimes also called area, areal, or grid data, are data that have been aggregated over some small geographic area. Often, a distinction is made between regular lattices (such as encountered for remote sensing data) or irregular lattices (such as states, counties, or health service areas). In a scenario where different economic regions are compared, a question of interest might be to explore how variables such as educational level, age, and racial composition of the population relate to unemployment in that region.
- c. Geostatistical data, sometimes also called spatially continuous data, are data that could, at least theoretically, be observed at any spatial location. However, cost and time determine at how many locations such data actually are collected. Examples of this type of data range from precipitation and temperature measurements to air pollution measurements and readings of minerals in the earth. A question of interest might be to visualize the distribution of nitrates in the soil in a specific region before fitting a smooth surface to the data.
- d. Origin–destination flow data, sometimes also called link or spatial interaction data, are data that consist of measurements, each of which is associated with a pair of point locations or a pair of areas. Examples for this type of data are home address and workplace address for inhabitants of a particular city or originating and destination airports for airline travel. A question of interest might be to explore from which originating airports most passengers, most flights, or most cargo arrives at a particular destination airport.

Cressie (1993) addressed the first three types of spatial data, both from a theoretical as well as from an applied perspective. When an additional temporal

component is available, that is, spatial data are collected over time; we speak of spatiotemporal data. Origin–destination flow data are discussed in detail in Fischer and Wang (2011, Part II). To a considerable extent, the underlying type of spatial data set determines which ESDA techniques are most suitable. Many of the ESDA techniques discussed in this chapter are suitable for more than one type of spatial data.

66.3 Basic Visualization and Exploration Techniques via Maps

The first step to explore spatial data often is to display the data on a map and then to produce several variations of the initial map. Credit needs to be given to John Snow (1813–1858), a British anesthesiologist, who was the first who mapped disease data. His investigation of the 1854 cholera outbreak in London pioneered the field of epidemiology. Nowadays, some consider him the “*father of epidemiology*” but the name “*grandfather of ESDA*” might suit him equally well. The 1854 London cholera outbreak started on August 19, 1854. It lasted about 6 weeks and resulted in more than 575 deaths. Snow (1936) observed: “... Mortality in this limited area probably equals any that was ever caused in this country, even by the plague.” Snow’s hypotheses were that cholera was transmitted from person to person via a fecal–oral route and that the drinking water of the Broad Street pump was the cause of the cholera outbreak. Snow utilized his map and empirical evidence to convince the Board of Guardians to remove the handle of the Broad Street pump. A mere 48 fatal attacks occurred, following the removal of the handle of the Broad Street pump, indicative that the water feeding the Broad Street pump could indeed be the source of the cholera epidemic.

As demonstrated by Snow, the visualization of spatial locations, that is, spatial point patterns, can provide valuable insights into such a data set. Moreover, if additional information is available for the locations such as age, gender, and case/control, this information can be displayed via different colors, symbols, and symbol sizes in the map display.

66.3.1 Choropleth Maps

For lattice data and geospatial data, several types of map displays exist and can be used for exploration. Best known, and most widely used, are choropleth maps. However, choropleth maps highly depend on choices made by the map creator. Even if the geographic boundaries are fixed as is the case for lattice data, Monmonier (1996, Chaps. 4 & 10) worked out different visual effects depending on whether the data are split into equal-interval classes or into quartile (or other quantile) classes. The same choices that affect histograms, that is, the starting point of a class interval and the width of each class interval, also affect choropleth maps.

Moreover, color choices in choropleth maps have a considerable effect on our perception. A small dark area in an overall bright map may (or may not) be

perceived as well as a small white area in an overall dark map. Excellent options for color choices for maps and statistical plots can be obtained from the *ColorBrewer* software tool (Harrower and Brewer 2003), accessible at <http://colorbrewer2.org>.

Finally, if geographic boundaries are not fixed in advance, choropleth maps can be easily affected by the modifiable areal unit problem (MAUP) (Openshaw 1984). Depending on the boundaries that are used for the aggregation (such as summation or averaging), rather different results for the sums, percentages, or averages may be obtained. Monmonier (1996, Chap. 10) demonstrated the MAUP for the locations of Snow's cholera data set. Therefore, it is necessary to explore what happens when spatial data get aggregated in different ways.

Given these different sources for biases when looking at choropleth maps, it is necessary to create and explore a variety of these maps to explore and understand the underlying spatial patterns. A single choropleth map that is the result of some default setting in a map-producing software package rarely will reveal all details of the underlying spatial data set. Andrienko et al. (2001) discussed how to conduct an exploratory analysis of spatial data via a combination of interactive maps and data mining.

66.3.2 Linked Micromap Plots

Linked micromap (LM) plots (Symanzik and Carr 2008; Carr and Pickle 2010) were introduced as an alternative to choropleth maps, especially to overcome some of the limitations of choropleth maps. The basic idea behind LM plots is to link geographic region names and their statistical values with their locations that are shown in a sequence of small maps, called micromaps. A typical LM plot (see Fig. 66.1) consists of three to five columns. The first column usually shows the maps, the second column lists some identifier (such as country or state names), and the third to the fifth columns contain statistical plots. Each small map highlights a few locations, typically five in a single map. The data are sorted according to some statistical criteria, for example, from highest to lowest (or vice versa), or from highest increase to lowest increase between years 1 and 2. Thus, the topmost map shows locations with the five largest (or smallest) observations according to the sorting criteria, the next map shows the five locations with the next largest (or smallest) observations, and so on. In case of any spatial association, locations with high or low observations tend to be plotted on the same map or on neighboring maps. The columns with the statistical plots may contain dot plots for each location, confidence intervals, time series plots, or box plots that are based on data for each particular location.

Micromaps have been used in print for applications as diverse as for comparisons of changing population density and population growth by state and for the visualization and interpretation of birth defects data in Utah and the United States. Typically, a published micromap is the result of many iterations where the authors experimented with different sortings and arrangements of the data panels and multiple possible layouts of the map panel.

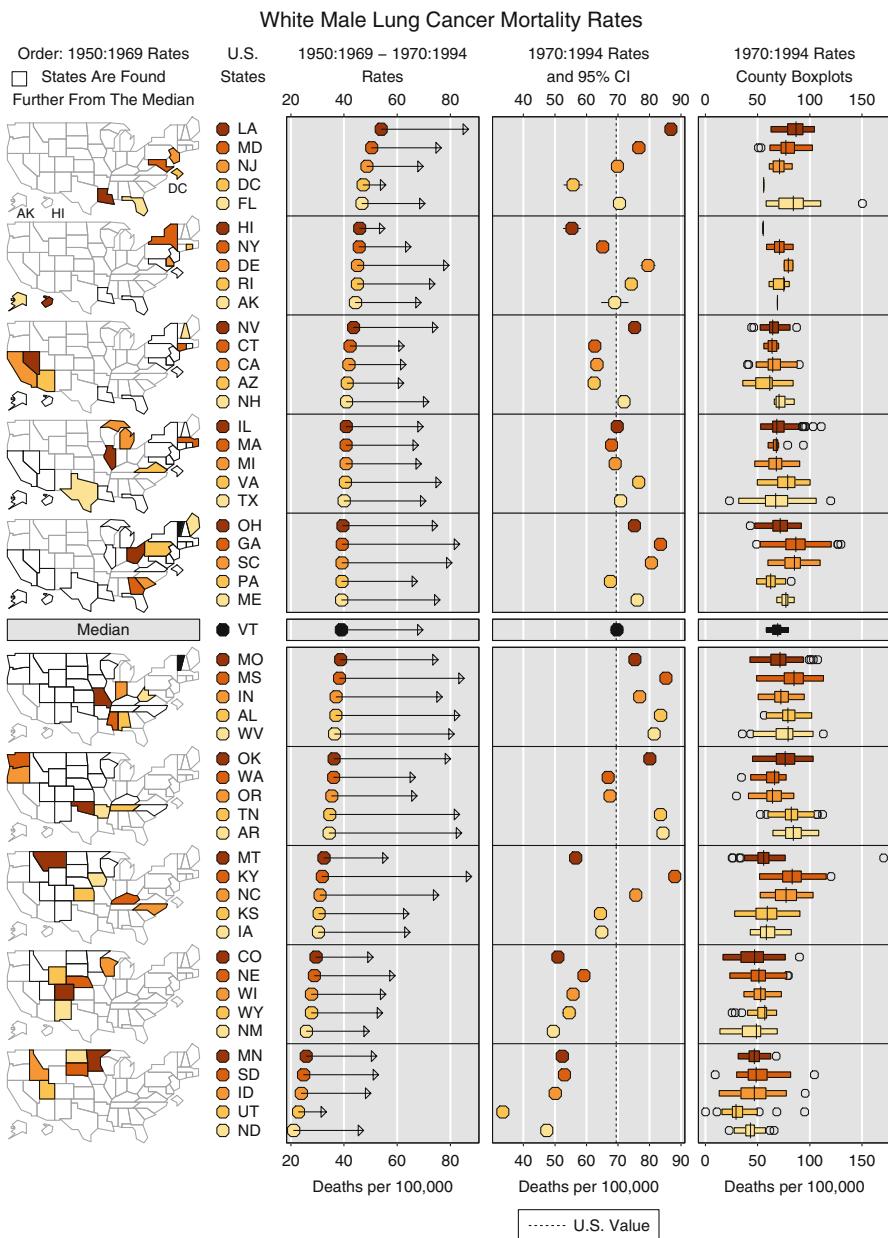


Fig. 66.1 LM plots, based on data from the NCI Web page, showing summary values for white male lung cancer mortality rates in the United States for the years 1950–1969 and for the years 1970–1994 in the *left data panel*, rates and 95 % confidence intervals in the *middle data panel*, and box plots for each of the counties of each state in the *right data panel* (Previously published as Fig. 1.6 in Symanzik and Carr (2008, p. 285))

While LM plots initially were constructed only for a static representation of the underlying data on paper or a computer screen, interactive versions may be introduced to allow an exploration of the underlying data from multiple perspectives. The US Department of Agriculture (USDA) – National Agricultural Statistics Service (NASS) Research and Development Division released an interactive micromap Web site (<http://www.nass.usda.gov/research/sumpant.htm>) in September 1999 for the display of data from the 1997 Census of Agriculture. This Web site still is accessible today. The National Cancer Institute (NCI) released an interactive micromap Web site (<http://www.statecancerprofiles.cancer.gov/micromaps>) in April 2003 for accessing their cancer data (Wang et al. 2002; Carr et al. 2002). This Web site is still accessible today and it is permanently updated with new data. While printed (static) LM plots are most suitable when the number of geographic regions ranges from about 10 to about 100, interactive LM plots may be suitable for several hundred geographic regions. Micromaps at the county level for the 254 counties of Texas at the NCI micromap Web site can reveal some very strong patterns, based on the data selection.

Figure 66.1 shows a static LM plot with three statistical columns based on data derived from the NCI Web site. The rows in the figure are sorted according to the 1950–1969 white male lung cancer rates in the United States (US) that reveal some strong geographic pattern with high rates in the eastern, southern, and western United States. A next step could be to resort the rows with respect to highest 1970–1994 rates, then with respect to highest absolute increases from the 1950–1969 to the 1970–1994 rates, and finally with respect to highest relative increases from the 1950–1969 to the 1970–1994 rates. Moreover, the second and third data column might be used to display data from possible confounding variables at the state level, such as smoking rates, gender composition, or educational level. After the exploration of several such LM plots, a researcher likely will have observed many known facts about the spatial distribution of male lung cancer, but, hopefully, some unexpected patterns and relationships also will have emerged.

66.3.3 Conditioned Choropleth Maps

Conditioned choropleth maps (CCmaps) (Carr et al. 2000; Carr and Pickle 2010) were introduced as a tool for the exploration of spatial data that consist of geographic locations, one dependent variable, and two independent variables. Via sliders, a researcher can interactively partition each of the two independent variables and the dependent variable into three different intervals each. A 3×3 set of panels containing nine partial maps shows the color-coded level (high, medium, low) of the dependent variable for those geographic locations that relate to high values of variable one and high values of variable two in map one, for those geographic locations that relate to high values of variable one and medium values of variable two in map two, and so on. For example, in an agricultural setting, variable one might be the amount of fertilizer, variable two might be the amount of

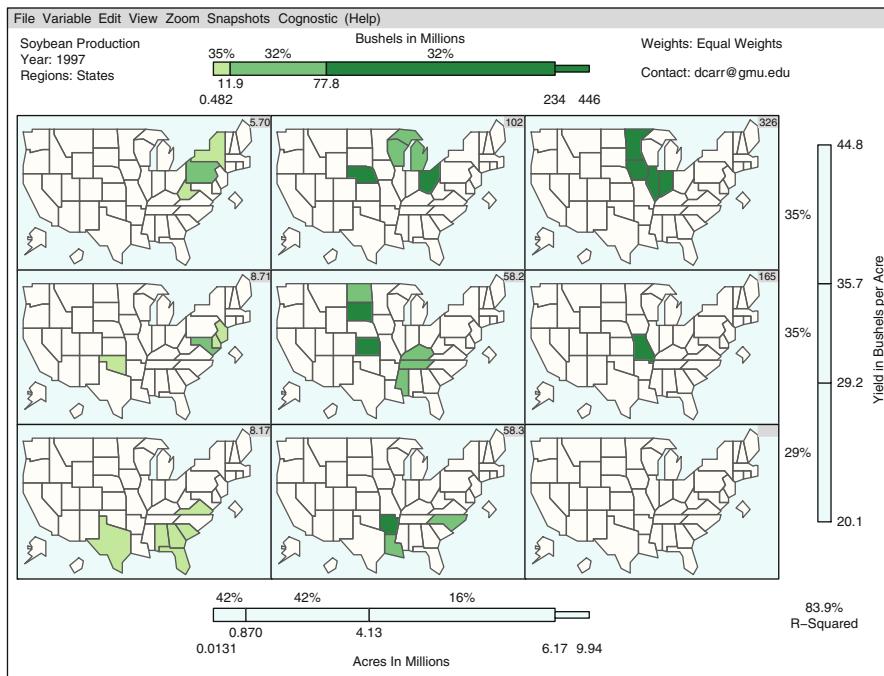


Fig. 66.2 CCmaps, based on data from the USDA–NASS Web page, related to soybean production in the United States. The plot shows the dependent variable production (*top slider*) that is conditioned on the two independent variables acreage (*bottom slider*) and yield (*right slider*) (Previously published as Fig. 1.7 in Symanzik and Carr (2008, p. 289))

precipitation, and the dependent variable might be the yield of a crop. One might expect that high values for fertilizer and precipitation result in a large yield. The nine maps show the relationship among the three variables in a geographic framework, thus allowing the consideration of the underlying spatial structure of the data and not only the statistical relationships. Cutoff values can be changed interactively, thus allowing the investigation of many possible settings. CCmaps are useful tools for the interactive generation of statistical hypotheses for medical, epidemiological, and environmental applications.

In Fig. 66.2, the 1997 soybean production in the United States is conditioned on acreage and yield. In an interactive environment, slider settings can be further modified to identify geographic areas of interest on the nine maps.

66.4 ESDA via Linking and Brushing

While in the previous section map views were interactively manipulated in a rather direct way, we will discuss in this section how map views and associated statistical

displays can be interactively manipulated via linked views and brushing. This is commonly understood as the classical idea of EDA and ESDA. For a detailed discussion of concepts for interactive graphics, such as brushing, linked brushing, linked views, focusing, zooming, panning, slicing, rescaling, reformatting, rotations, projections, and the grand tour, the reader is referred to Symanzik (2004, Sect. 10.3). Main statistical plot types that can be frequently found as components of linked views include histograms, scatterplots, scatterplot matrices, the grand tour, parallel coordinate plots, bar charts, pie charts, spine plots, mosaic plots, ray-glyph plots, and cumulative curves (such as the Lorenz curve). Most of these plot types also were discussed in Symanzik (2004, Sect. 10.3). Figure 66.3 shows one map view that has been linked with two scatterplots. In addition, plots for spatial data, such as variogram–cloud plots (see Fig. 66.4) and spatially lagged scatterplots, can be components of the linked views.

The overall idea of brushing is to mark different subsets of the data in a particular plot with different colors, symbols, sizes, or point or line styles. This is usually done based on the visual appearance of patterns in a specific plot, for example, outliers that seem to be far away from the remaining points in a histogram or scatterplot, or clusters that seem to be well separated from each other. In the next plot that is being produced, the original assessment will be reevaluated, additional points may be marked, or points may be marked differently.

In the framework of linked brushing and linked views, the brushing information is carried over from one plot to the next. For example, outliers that are marked in a histogram or scatterplot will be marked in a similar way (with the same colors, symbols, sizes, or point or line styles) in all related plots, in particular on a map view as well. Monmonier (1989) introduced the term geographic brushing in reference to interacting with the map view of geographically referenced data.

In Fig. 66.3, the US cities with the highest index for education have been brushed in the left scatterplot, and cities with a high crime index have been identified by name in the same scatterplot. The map view shows the locations of these cities with the same color and symbols as in the scatterplot. Moreover, the scatterplot on the right reveals that a high crime index is associated with a high recreation index while a high education index is associated with a medium recreation index. Nothing striking is noticeable when comparing the brushed values for education and crime with the arts index. Extensions of brushing for spatial data, such as moving statistics, or brushing, applied to origin–destination flow data (Liu and Marble 1997), exist.

In advanced software environments, brushing can take place in any of the linked views, including the map view. So, when locations in a specific geographic region are marked, the other statistical views will reveal whether there is some possible statistical relationship among the data from the selected locations as well, for example, whether the statistical values are similar to each other or whether the statistical values span the entire range of the underlying data distribution.

Linked brushing is not always one-to-one between the different displays. In a variogram–cloud plot, the absolute difference (or a related measure) of a variable of interest is calculated for all pairs of spatial locations, and this measure

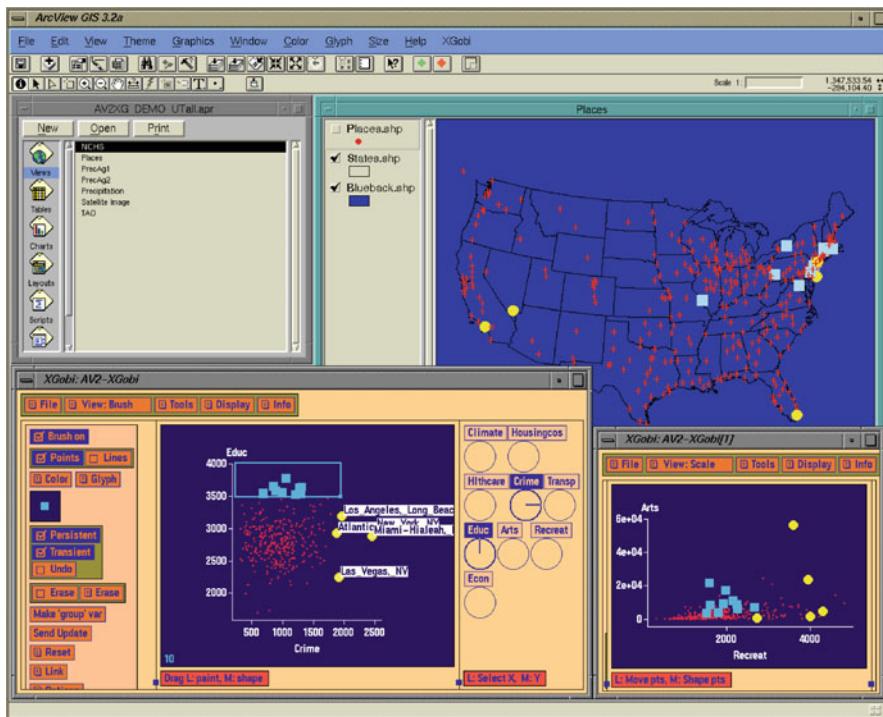


Fig. 66.3 Screenshot of the “Places” data in ArcView/XGobi. A map view of 329 cities in the United States is displayed in ArcView at the top. The two XGobi windows at the bottom are showing scatterplots of crime (horizontal) versus education (vertical) (left) and recreation (horizontal) versus arts (vertical) (right). Locations of high crime have been brushed and identified, representing some of the big cities in the United States. Also, locations of high education (above 3,500) have been brushed, mostly representing locations in the northeastern United States. All displays have been linked (Previously published as Fig. 10.1 in Symanzik (2004, p. 299))

is plotted against the Euclidean distance between the two associated points (up to a cutoff distance chosen by the researcher). Thus, when brushing one point in a variogram–cloud plot, this needs to be translated to a pair of spatial locations that are brushed in the map view.

Figure 66.4 shows such a link for precipitation measurements in the northeastern United States. In this figure, the highest values in the variogram–cloud plot (up to the cutoff distance) have been brushed. The map view shows two points, that is, spatial locations, that are connected to several other spatial locations. The location in the northeast likely is a spatial outlier as its precipitation measurements are considerably different (either higher or lower) than those from all nearby locations. A next step would be to explore additional variables for these locations, starting with elevation. The location in the southwest likely is not a spatial outlier; rather, there is some considerable local variation happening in this region as this location is only connected to some, but by far not all, locations in its neighborhood.

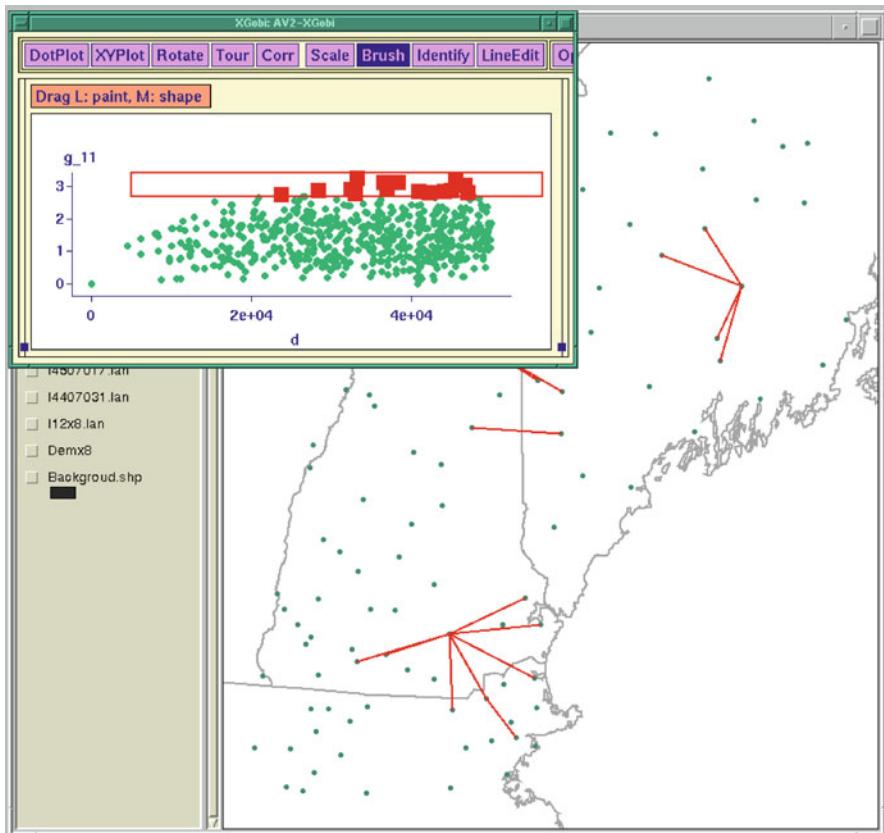


Fig. 66.4 Example of a variogram–cloud plot that is linked to a map view, based on precipitation data for the northeastern United States. In the *upper left* XGobi window, we have brushed (using a *solid rectangle*) the highest values in the variogram–cloud plot. In the *lower right* ArcView map view, each pair of locations, related to a point that has been brushed in XGobi, has been connected by a line (Previously published as Fig. 2 in Symanzik et al. (2000, p. 477). Reprinted with permission from the *Journal of Computational and Graphical Statistics*. Copyright 2000 by the American Statistical Association. All rights reserved)

66.5 Local Indicators of Spatial Association (LISA)

Moran's I statistic is a well-known measure for spatial autocorrelation at the global level for lattice data. Anselin (1995) introduced a local Moran statistic, a local Gamma statistic, a local Geary statistic, a Moran scatterplot, and other LISA statistics to assess the spatial association at a location i . The LISA statistics allow to identify local spatial clusters and to assess local instability. Moreover, the LISA statistics allow to assess the influence of a single location on the corresponding global statistic, a feature that is as important as being able to identify influential points in a regression framework.

LISA statistics are probably the most frequently applied ESDA technique, with applications in areas as diverse as regional sciences, spatial econometrics, epidemiology, social sciences, and criminology. Despite its wide use, one should keep in mind that LISA statistics are exploratory in nature, and, usually, additional steps are required to confirm the initial results derived from LISA statistics.

66.6 Software for ESDA

ESDA is highly dependent on software that supports various types of statistical displays, map views, and that allows linked brushing. Two main approaches have been developed during the last 25 years: Conducting ESDA in software environments where a geographic information system (GIS) is linked to statistical software packages and in stand-alone statistical software solutions. A more detailed overview of various software solutions for ESDA has been provided in Symanzik ([2004](#), Sect. 10.6.1).

66.6.1 ESDA and GIS

Fotheringham ([1992](#)) pointed out that it is not necessary to conduct an exploratory spatial data analysis within a GIS, but that in many circumstances, using a GIS to do so might simplify the exploration of the data and provide insights that otherwise might be missed. Therefore, over the next decade, several researchers developed software that linked GIS with statistical software, or they added statistical features to existing GIS.

In Anselin ([1994](#)), a series of ESDA techniques were discussed in the context of a GIS, with the primary focus on exploring the spatial nature of the underlying data. These techniques could be classified as techniques based on the neighborhood view of spatial association (such as Moran scatterplots and LISA statistics) and as techniques based on the distance view of spatial association (such as spatially lagged scatterplots and variogram–cloud plots). Various software links between GIS such as Arc/Info, ArcView, and Grassland and one or more statistical software packages implemented several of these techniques. Some of the links that were developed and maintained over a longer time period were links between Arc/Info, respectively, ArcView, and SpaceStat (Anselin et al. [1993](#); Bao and Anselin [1997](#)) and links between ArcView, XGobi, and XploRe (Cook et al. [1996](#); Symanzik et al. [2000](#)). One major limitation of such software links is that whenever one of the individual software packages is modified with respect to the functionality of the link, the other software packages have to be modified accordingly.

66.6.2 Stand-Alone Software for ESDA

In contrast to linking GIS and statistical software, several software developers focused on the development of stand-alone statistical software that also support map views of the spatial locations that are linked with statistical displays. Some of

the best known examples are Spider (Haslett et al. 1990), REGARD (Unwin et al. 1990; Unwin 1994), and, more recently, GeoDa (Anselin et al. 2006). One major limitation of stand-alone software for ESDA is that the functionality that is usually available in a GIS has to be reimplemented in a statistical software package.

In recent years, R (R Development Core Team 2011) has become the *lingua franca* of statistics. Since its appearance around 1996 (Ihaka and Gentleman 1996), R has been further advanced by thousands of creators of contributed packages (almost 4,000 in May 2012) that provide all kinds of additional functionality beyond the original R base functionality. This includes packages for maps, color selections, EDA and ESDA, and advanced statistical functionality for spatial data, such as the following:

- *maptools* (<http://cran.r-project.org/web/packages/maptools/index.html>) that allows to read and manipulate geographic data, in particular ESRI shapefiles
- *maps* (<http://cran.r-project.org/web/packages/maps/index.html>) that provides access to a variety of maps
- *RgoogleMaps* (<http://cran.r-project.org/web/packages/RgoogleMaps/index.html>) that allows to query the Google server for static maps and to use one of the Google maps as a background image to overlay statistical plots from within R
- *RColorBrewer* (<http://cran.r-project.org/web/packages/RColorBrewer/index.html>), the R implementation of <http://colorbrewer2.org>, for good color choices for maps and other plots
- *iplots* (<http://cran.r-project.org/web/packages/iplots/index.html>), an R package in the spirit of Spider and REGARD, for interactive plots in R, including maps
- *splancs* (<http://cran.r-project.org/web/packages/splancs/index.html>) for the exploration and analysis of spatial and space–time point patterns,
- *spatstat* (<http://cran.r-project.org/web/packages/spatstat/index.html>) for the exploration and analysis of spatial data, mainly spatial point patterns
- *spdep* (<http://cran.r-project.org/web/packages/spdep/index.html>) for the analysis of spatial dependence at a local and global scale, including Moran and LISA statistics
- *geoR* (<http://cran.r-project.org/web/packages/geoR/index.html>) for the exploration and analysis of geostatistical data
- *gstat* (<http://cran.r-project.org/web/packages/gstat/index.html>) for modeling, prediction, and simulation of spatial and spatiotemporal geostatistical data
- *spgrw* (<http://cran.r-project.org/web/packages/spgwr/index.html>) for computing geographically weighted regression

While the Web pages listed above provide detailed user guides and information how to use each of these packages, Bivand (2010) demonstrated how many of the ESDA techniques described in this chapter can be performed in R. An extended overview of additional R packages for the reading, exploration, visualization, and analysis of spatial data can be found at <http://cran.r-project.org/web/views/Spatial.html>.

66.7 Conclusions

In this chapter, we have provided an overview of techniques, methods, and software solutions for ESDA. Most of the developments took place during the

last 25–30 years. Due to the rapid development of computer hardware, including high-quality graphic displays, over the last few decades, ESDA techniques are nowadays easily accessible for many researchers on a wide variety of hardware platforms.

A current hotspot for ongoing development of ESDA techniques is R and its thousands of contributed packages. For a few decades, software packages for exploratory data analysis were relatively weak for confirmatory data analysis (using John Tukey's terms here), and vice versa. However, R is continuously getting stronger for both types of data analyses, and it is able to handle a large variety of GIS data formats. It can be expected that in the near future, exploratory and confirmatory data analyses will be conducted almost simultaneously in R or some similar software environment. Once a researcher detects something of interest in a spatial data set via ESDA, a confirmatory analysis can immediately follow, and once a confirmatory analysis has been conducted, ESDA can be used to further explore the spatial fit of the fitted model, its residuals, and so on.

A trend in recent years has been to provide access to spatial data for everyone via Web interfaces. This includes the previously introduced Web sites for interactive micromaps (<http://www.nass.usda.gov/research/sumpant.htm> and <http://www.statecancerprofiles.cancer.gov/micromaps>), but, even more, Web-based software such as *gapminder* (Rosling and Johansson 2009), accessible at <http://www.gapminder.org/>. The Google version, called *Google Public Data Explorer*, accessible at <http://www.google.com/publicdata/directory>, might become a tool that provides easy and fast access to EDA and ESDA techniques for millions of Web users.

66.8 Cross-References

- [Spatial Clustering and Autocorrelation in Health Events](#)

References

- Andrienko N, Andrienko G, Savinov A, Voss H, Wettschereck D (2001) Exploratory analysis of spatial data using interactive maps and data mining. *Cartogr Geogr Inform Sci* 28(3):151–165
- Anscombe FJ (1973) Graphs in statistical analysis. *Am Statistician* 27(1):17–21
- Anselin L (1994) Exploratory spatial data analysis and geographic information systems. In: Painho M (ed) New tools for spatial analysis. Eurostat, Luxembourg, pp 45–54
- Anselin L (1995) Local indicators of spatial association – LISA. *Geogr Anal* 27(2):93–115
- Anselin L, Dodson RF, Hudak S (1993) Linking GIS and spatial data analysis in practice. *Geogr Sys* 1(1):3–23
- Anselin L, Syabri I, Kho Y (2006) GeoDa: an introduction to spatial data analysis. *Geogr Anal* 38(1):5–22

- Bao S, Anselin L (1997) Linking spatial statistics with GIS: operational issues in the SpaceStat–ArcView link and the S + Grassland link. In: 1997 proceedings of the section on statistical graphics. American Statistical Association, Alexandria, pp 61–66
- Bivand RS (2010) Exploratory spatial data analysis. In: Fischer MM, Getis A (eds) Handbook of applied spatial analysis: software tools, methods and applications. Springer, Berlin/Heidelberg, pp 219–254
- Carr DB, Pickle LW (2010) Visualizing data patterns with micromaps. Chapman & Hall/CRC, Boca Raton
- Carr DB, Wallin JF, Carr DA (2000) Two new templates for epidemiology applications: linked micromap plots and conditioned choropleth maps. *Stat Med* 19(17–18): 2521–2538
- Carr DB, Chen J, Bell BS, Pickle LW, Zhang Y (2002) Interactive linked micromap plots and dynamically conditioned choropleth maps. In: Proceedings of the second national conference on digital government research, Digital Government Research Center (DGRC), pp 61–67, http://www.dgrc.org/conferences/2002_proceedings.jsp
- Cook D, Majure JJ, Symanzik J, Cressie N (1996) Dynamic graphics in a GIS: exploring and analyzing multivariate spatial data using linked software. *Comput Stat* 11(4):467–480. Special issue on computeraided analysis of spatial data
- Cressie NAC (1993) Statistics for spatial data, revised edn. Wiley, New York
- Fischer MM, Wang J (2011) Spatial data analysis:models, methods and techniques. Springer, Berlin/Heidelberg/New York
- Fotheringham AS (1992) Exploratory spatial data analysis and GIS. *Environ Plann A* 24(2):1675–1678
- Harrower MA, Brewer CA (2003) ColorBrewer.org: an online tool for selecting color schemes for maps. *Cartogr J* 40(1):27–37
- Haslett J, Wills G, Unwin A (1990) SPIDER – an interactive statistical tool for the analysis of spatially distributed data. *Int J Geogr Inform Syst* 4(3):285–296
- Ihaka R, Gentleman R (1996) R: a language for data analysis and graphics. *J Comput Graph Stat* 5(3):299–314
- Liu L, Marble D (1997) Brushing spatial flow data sets. In: 1997 proceedings of the section on statistical graphics, American Statistical Association, Alexandria, pp 67–72
- Monmonier M (1989) Geographic brushing: enhancing exploratory analysis of the scatterplot matrix. *Geogr Anal* 21(1):81–84
- Monmonier M (1996) How to lie with maps, 2nd edn. University of Chicago Press, Chicago
- Openshaw S (1984) The modifiable areal unit problem. In: Concepts and techniques in modern geography No. 38. Geo Books, Regency House, Norwich
- R Development Core Team (2011) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3–900051–07–0. <http://www.R-project.org/>
- Ripley BD (1981) Spatial statistics. Wiley, New York
- Rosling H, Johansson C (2009) Gapminder: liberating the x-axis from the burden of time. *Stat Comput Stat Graph Newslett* 20(1):4–7
- Snow J (1936) Snow on cholera: being a reprint of two papers by John Snow, M.D. together with a biographical memoir by B. W. Richardson, M.D. and an introduction by Wade Hampton Frost, M.D. The Commonwealth Fund/Oxford University Press, New York/London
- Symanzik J (2004) Interactive and dynamic graphics. In: Gentle JE, Härdle W, Mori Y (eds) Handbook of computational statistics – concepts and methods. Springer, Berlin/Heidelberg, pp 293–336
- Symanzik J, Carr DB (2008) Interactive linked micromap plots for the display of geographically referenced statistical data. In: Chen C, Härdle W, Unwin A (eds) Handbook of data visualization. Springer, Berlin/Heidelberg, pp 267–294 & 2 color plates

- Symanzik J, Cook D, Lewin-Koh N, Majure JJ, Megretskiaia I (2000) Linking ArcView and XGobi: insight behind the front end. *J Comput Graph Stat* 9(3):470–490
- Tukey JW (1977) Exploratory data analysis. Addison Wesley, Reading
- Unwin A (1994) REGARDing geographic data. In: Dirschedl P, Ostermann R (eds) Computational statistics. Physica-Verlag, Heidelberg, pp 315–326
- Unwin A, Wills G, Haslett J (1990) REGARD – graphical analysis of regional data. In: 1990 proceedings of the section on statistical graphics, American Statistical Association, Alexandria, pp 36–41
- Wang X, Chen JX, Carr DB, Bell BS, Pickle LW (2002) Geographic statistics visualization: web-based linked micromap plots. *Comput Sci Eng* 4(3):90–94

Spatial Clustering and Autocorrelation in Health Events

67

Geoffrey Jacquez

Contents

67.1	Introduction	1311
67.2	Background and Trends	1312
67.3	Scientific Inference from Patterns of Health Events	1314
67.4	Strong Inference for Health Events	1315
67.5	Sources of Spatial Autocorrelation in Health Events	1315
67.6	Role of Neutral Models	1325
67.7	Data: What to Cluster?	1326
67.8	Data Models and Clustering Methods	1328
67.9	Conclusions	1332
	References	1333

Abstract

Spatial autocorrelation in health events may be the signature of underlying causal factors of direct scientific and practical interest but may also be due to pedestrian or nuisance factors that obscure meaningful spatial patterns. The problem is to discern spatial patterns that inform our understanding of the health events themselves from those that are of little interest. This chapter provides a framework for advancing knowledge when the causes of observed health event clusters are unknown.

G. Jacquez
SUNY at Buffalo, Buffalo, NY, USA

BioMedware, Ann Arbor, MI, USA
e-mail: geoffrey.jacquez@biomedware.com

67.1 Introduction

We begin with background and trends in health data, electronic health and mobile health, and clustering as pre- versus post-epidemiology. Motivations for clustering are presented, with the scope of inference and what can be learned from spatial disease patterns. The approach of Strong inference is described and proposed as a useful framework for the analysis of health event clusters. This touches on explanations for patterns, which can be due to nuisance factors and covariates, as well as to underlying causes of disease. Coverage is then given to sources of autocorrelation in health events, which are extensive, and range from the signature of underlying disease processes to covariates and risk factors that may or may not be of interest, to treatment effects, access to care, and geographic variation in pathogen and host genetics, as well as other causes. Each of these may be an explanatory hypothesis describing the origin of a health event cluster. Complicating factors include latency and temporal lag processes, location uncertainty, and using location as a proxy. Next, the types of data that may be analyzed are presented, ranging from case locations, case-control data, mobility histories, and data aggregated into rates. What goes into the denominator is an important consideration when using rates, and alternatives, such as the stage of diagnosis that have readily estimated denominators, are considered. Consideration is given to the kinds of data that may be analyzed, including incident cases and deaths, case-control data, symptoms, volunteer geographic information and crowd sourcing, and residuals from spatial models. Examples of data models for these different types are next. The conclusion revisits important problems and future prospects, including sources of false negatives in cluster analyses, the use of location as a surrogate for exposure or as a proxy for another variable, and the development of experimental methods and sampling frameworks for the emerging era of “big data.” “Big data” refers to massive data sets that collect over time and that are difficult to analyze using common database management tools; these pose a true challenge in spatial analysis. Very large data sets have been around for quite some time; what distinguishes “big data” is the heterogeneity of its sources that include retail transactions, photos and surveillance videos, data from logs and sensors, as well as unstructured text posted on the Web, such as blogs and social media.

67.2 Background and Trends

Once identified, health event clusters may be used to guide public health response, to site clinics and screening facilities in order to better serve at-risk communities, to guide application of health interventions, and to formulate health policies. The accurate and timely identification of spatial and temporal patterns in health events is therefore of some importance (Kingsley et al. 2007).

Pre- Versus Post-Epidemiology. Health event clustering has been referred to as “pre-epidemiology” since it has often relied on encountered data that were not

collected using an experimental design suited to cluster detection. This has limited its inferential power to the extent that clustering has been viewed as an expensive activity that does little to increase our understanding of the causes of disease (Neutra 1990). As the field has advanced, clustering has been applied to data that come from traditional experimental designs such as case-control studies, and techniques have been developed for systematically excluding hypotheses that might explain observed disease patterns. One then seeks to allocate the risk not explained by the known risk factors to specific places, groups of study participants, and times. This supports the construction of new hypotheses regarding underlying causal factors and may be thought of as “post-epidemiology” as the analyses are conducted after, or as an adjunct to, traditional epidemiological analyses. The key is that the data being analyzed have been collected as part of a sampling design that controls for known risk factors and covariates.

Application Areas. Initially, health event cluster analysis was primarily a response mechanism for replying to cluster alarms raised by a concerned public. This was a retrospective analysis in that the techniques were applied to existing data to evaluate whether there was a statistically significant excess of disease. Surveillance is used when one wishes to analyze a stream of health data in order to detect an increase or change in baseline risk. Syndromic surveillance is a special case of surveillance that uses health indicators that may not have been fully vetted via diagnoses, laboratory confirmation, or other means. Here the objective is the timely identification of disease outbreaks, such as might arise under a bioterrorist attack or from local emergence of a novel flu strain. Syndromic surveillance has evolved as data on indicators of disease, such as pharmaceutical prescriptions and over-the-counter sales of cough, diarrhea, and other medications, have become increasingly available. Meaningful use guidelines for syndromic surveillance in public health using electronic health record data collected in emergency department and urgent care settings are being proposed at the time of this writing (Johnson et al. 2012). As noted above, techniques for analyzing volunteered geographic information on disease are now being developed. With the advent of mobile devices, such as cell phones, mobility traces coupled with information on health outcomes may be clustered to assess relationships between mobility patterns, specific places, and health events.

The Changing Landscape of Public Health Data. The availability of georeferenced data in health analysis is expanding rapidly due to several technological and policy trends. First, there is increased availability of user-generated, location-enabled health data as segments of the population become comfortable with sharing information through smart phones, Web browsers, and other means and as search engine keywords and social media are used to assess near real-time trends in health-related symptoms, medications, and outcomes. The confluence of crowd sourcing (e.g., “reflexive consumerism” where patients review hospitals and professionals on the Web) and volunteer geographic information (VGI, where individuals report activities at their location) is enabling significant advances in disaster response, epidemiology, and exposure assessment science (Goodchild and

Alan Glennona [2010](#)). For example, by coupling technologies for near real-time sensing of pollutants with location-enabled devices such as mobile phones, VGI can be used to generate model-based high spatial resolution exposure estimates. This makes possible validation of individual-level exposure estimates as a person goes about their daily activities.

Second, the US health-care system and the Department of Health and Human Services are investing heavily in interoperable electronic health records expected to revolutionize health care and disease control and surveillance. Recent national legislations such as the Health Information Technology for Economic and Clinical Health (HITECH) Act and the Affordable Care Act (ACA) include provisions requiring the collection of detailed electronic data in standardized format for insurance and care equity purposes ([Weissman and Hasnain-Wynia 2011](#)). Many of the data records for these systems include personal identifiers – names, addresses, and related health information – that can be used to construct georeferenced databases on patients, providers, and health-related resources such as screening facilities.

Third, advances in spatiotemporal epidemiology facilitate reconstruction of geocoded residential histories of patients. The feasibility of developing reliable geospatial data retrospectively for large, epidemiological studies has been demonstrated, and revisiting completed studies using spatial epidemiological methods is now possible. In an era of fiscal constraints, expensive, large epidemiological studies are less likely to be funded. Application of spatiotemporal analysis to completed case-control, cohort, and longitudinal studies holds enormous promise for gaining new insights into disease causation that leverages existing investments in health research.

67.3 Scientific Inference from Patterns of Health Events

Health event clusters may loosely be defined as statistically significant excesses of health events in space, in time, or in space-time. There also is space-time interaction, as when nearby health events occur at about the same time. Cluster existence, location, and timing can inform decisions regarding different questions, such as:

1. Is an observed pattern of health events statistically unusual? (Is apparent clustering real?)
2. Where are populations with elevated disease rates? (Where are local excesses found?)
3. Are areas with elevated health events found in proximity to geographic features thought to be associated with disease causality? (Is there focused clustering about pollutant sources?)
4. Is the observed spatial pattern of health events consistent with certain hypothesized disease processes, and not consistent with others (what is the underlying cause?)
5. Are there reasonable new hypotheses that might explain the observed disease patterns (what is the best explanation for the cluster)?

Several of these questions can be addressed using an inferential process where plausible generating processes for an observed pattern are considered and then excluded. This can be done in a haphazard fashion, but it usually is best to systematically enumerate the set of plausible hypotheses that might give rise to an observed pattern of health events and to then exclude members of this set by conducting a series of experiments that may include statistical tests and models for evaluating space-time disease patterns. This inferential framework seeks to accomplish a mapping of health event patterns to the spatial processes that might give rise to them, and is called Strong inference.

67.4 Strong Inference for Health Events

In 1964 Platt coined the term “Strong inference” (Platt 1964) to describe a useful construct for systematically evaluating explanatory hypotheses that plausibly might explain observed patterns in a data set. It involves, first, enumeration of the explanatory hypotheses that might give rise to the pattern; second, formulation of falsifiable predictions that can be used to systematically test each of these hypotheses; third, undertaking the tests of predictions; and, fourth, winnowing out the hypotheses whose corresponding predictions are found to be false. The remaining hypotheses then must include, or together explain, the observed data patterns. The initial set of explanatory hypotheses may be expanded as the experiments are conducted. What is key is that the predictions framed for each hypothesis be falsifiable (e.g., can be tested using a statistic for spatial clustering) and that the set of explanatory hypotheses be properly framed.

67.5 Sources of Spatial Autocorrelation in Health Events

Spatial autocorrelation is characteristic of almost all geographic data and can reflect the magnitude and spatial scale of underlying causal processes (Getis 2010). This raises a very important question. What are the sources of spatial autocorrelation in health events? These may need to be included in the set of explanatory hypotheses for an observed pattern and include spatial autocorrelation in underlying risk factors, covariates, reporting, diagnosis, health-care policies, physician behaviors, and interpolation autocorrelation, as summarized below. This is by no means an exhaustive list but includes factors that likely should be considered in many spatial analyses of health events.

Multifactorial Causes of Disease. It is important to recognize that many health outcomes may be caused by several different disease processes and that a given exposure mechanism may result in different disease outcomes. For example, risk factors for myocardial infarction include genetic predispositions, diet, body weight, exercise habits, medication compliance, and access to care, among others. And specific exposures, such as smoking, are associated with elevated risk for a host of health outcomes, including bladder, throat, and lung cancers, asthma, pneumonia, and emphysema, among others.

Spatial autocorrelation in health events may arise whenever the host of factors underlying disease expression are themselves spatially structured. Genetic predispositions for disease may be inherited, giving rise to spatial autocorrelation in disease risk whenever family members cohabit and tend to live near one another; ambient air pollutant concentrations tend to be highly spatially autocorrelated; and so on.

Comorbidity and Competing Causes of Death. It is unusual for chronic diseases to be the sole disease process occurring in a patient, especially as the age of the subject increases. This makes sense when one considers the multifactorial nature of most diseases. At the population level, smoking will increase the risk of both lung and bladder cancers; at the individual level, a smoker may have comorbid conditions such as emphysema and lung cancer. The expression of infection processes is often mediated by immune response and the health status of the individual. Hence, risk of infection increases as the physical condition of the individual declines. Individuals and populations are thus both subject to competing causes of death. Prior to the advent of antibiotics in the 1940s, respiratory and childhood infections were major sources of mortality in most developed countries. As antibiotics became widely available, the major source of mortality became chronic diseases such as heart conditions and cancer. These were “unmasked” once respiratory and childhood infections were removed as a competing cause of death. Spatial autocorrelation in health events thus may arise when there is underlying geographic variation in comorbid conditions and/or risks for competing causes of death.

Geographic Variation in Exposure and Behaviors That Mediate Exposure. Health events associated with environmental exposures are mediated by exposure routes including eating, drinking, breathing, dermal exposures, and ionizing radiation. When considering health outcomes associated with exposure to specific risk factors, such as arsenic, one needs to consider relevant exposure routes and mechanisms, such as consumption of foods and beverages containing biologically active forms of arsenic. Exposure-mediating behaviors are often modifiable risk factors, since what one smokes, drinks, and eats are to a certain extent individual choices that can be changed. When evaluating spatial patterns in health outcomes associated with environmental exposures, one needs to consider both environmental concentrations as well as the exposure routes whereby the compound under consideration enters the body. Both of these (environmental concentrations and exposure routes and mechanisms) may themselves be spatially structured. For example, the amount of water people drink varies with age, decreasing as one gets older; with occupation (farm workers requiring more water than office workers); with altitude; and other factors.

Socioeconomic and Demographic Factors. One definition of “covariate” is a variable that has an effect (e.g., is associated with the outcome) that is not of direct interest. When modeling health events such as disease incidence, socioeconomic and demographic factors such as age may be considered as covariates, since age, for example, does not of itself cause disease. Yet, these are of considerable importance when evaluating spatial disease patterns, since the risk of most health outcomes including cancer, heart disease, and infections is typically associated with socioeconomic status, sex, race, and age. One thus may need to account for spatial

patterns in covariates when assessing the significance of a clustering of health events. Rather than asking “Are the health events clustered?” one instead may ask “Is there significant clustering of health events above and beyond spatial patterns in covariates?” Neutral models (described later) have been developed to address this question.

Genetics. Microevolutionary processes such as selection, isolation by distance, and migration give rise to spatial autocorrelation in genetic structure and genetic variance in geographically distributed populations. While we often think of human populations as being very well mixed, interbreeding freely over large geographic distances, this is often not the case. Population genetics in North American, European, and Asian populations have been demonstrated to be spatially autocorrelated and associated with language and dialect. This makes sense when one considers that children speak the language of their parents and family and that family members tend to live in geographic proximity of one another, even though some may travel far from their homes. Familial clusters are often observed for many cancers, both because of behavioral factors mediating common exposures such as secondhand smoking and diet but also because of within-family genetic similarity in oncogenes and tumor suppressor genes. For example, one hypothesis for explaining the excess of breast cancer incidence on Long Island is the higher incidence of mutations of BRCA genes in local populations thought to be descended from European populations. BRCA1 and BRCA2 are tumor suppressor genes, and mutations of these genes have been linked to breast and ovarian cancers.

For infectious diseases the pathogen, whether a virus or a bacteria, undergoes a population bottleneck whenever there is infection transmission to a susceptible person. Only a few (e.g., several thousands) of the pathogen may be required for infection to take hold, and together these may have a genetic composition that is quite different from the overall pathogen population. A mutation that occurs during a bottleneck can become fixed in the pathogen population infecting that host (person). When this mutation is associated with changes in infection transmission or severity of the infection, it can have important consequences for the spread of infection, as well as for morbidity, mortality, and resistance to treatment. Such mutations can give rise to new pathogen strains, and the occurrence of these strains may be observed as outbreaks of the new strain, initially occurring in localized populations. This has been documented for diverse infectious diseases including cholera, tuberculosis, HIV, and influenza.

Perhaps one of the best known instances of interactions between infection and genetics is selective pressures for the sickle cell trait that foster resistance to malaria infection. In sickle cell disease, the red blood cells are misshapen, leading to circulatory problems and early death of red blood cells, resulting in anemia. The disease has a genetic basis, with alleles that code for the sickle cell trait and for abnormal hemoglobin, resulting in different forms of the disease of varying severity. But when one sickle cell allele is present, it confers some resistance to malaria infection. This confers a substantial selective pressure in populations residing in malarial regions. The sickle cell trait and sickle cell anemia, thus, vary

geographically with higher penetration of the sickle cell gene in populations residing where malaria is endemic.

Vector-borne diseases and parasites often have complex life histories, involving infection transmission and amplification among humans and one or more host organisms. Well-known examples include malaria, Lyme disease, and West Nile virus, among others. Here, spatial structure in the genetics of the pathogen can arise due to the interactions between population bottlenecks and mutations, as noted above for infectious diseases. The genetics of the host species can also influence the origin and spread of different pathogen strains.

Environment/Vector-Pathogen Ecology. Environmental patchiness in habitats suitable for vector and host organism survival is an important determinant of where and when vector-borne and parasitic infections occur. In the northeastern and midwestern United States, the white-tailed deer (*Odocoileus virginianus*) is an important host species for Lyme disease, which is transmitted by a bite from infected blacklegged ticks. Infection transmission events can only occur where both infected ticks and susceptible people are present. Blacklegged tick habitat includes wooded, brushy areas that provide food and cover for intermediate host species such as white-footed mice and white-tailed deer. But infection transmission to humans only occurs when people are in areas where infected ticks are present and feeding. Thus, the occurrence of Lyme disease is highly associated with geographic overlap of human activity spaces with habitat suitable for both intermediate hosts and the tick itself. Infection transmission is highly structured temporally as well, occurring in those months when the tick is searching for blood meals in the spring and fall.

Heterogeneity in Population Density, Rate Stability, and the Small Numbers Problem. Health events that occur in small areas may be expressed as a rate, such as an incidence or mortality rate. Rates are calculated from a numerator, such as the number of incident lung cancer cases in white males, and a denominator, such as the population at risk (e.g., white males) for lung cancer. The rate is calculated by dividing the numerator by the denominator, and this is where the “small numbers problem” arises. The variance in the rate depends critically on the size of the denominator. When the denominator is small, variance in the rate is high; when the denominator is large, variance in the rate is small. Hence, the appearance of an apparently large rate might be due entirely or in part to the small numbers problem (e.g., a small denominator with a resulting large variance in the rate estimate), and the true, underlying risk might be entirely unremarkable. A simple protocol for evaluating whether the small numbers problem is having an impact on estimated rates is as follows. First, create a map of the rate and a scatterplot of the rate (on the x-axis) and the population at risk (on the y-axis). Next, inspect the scatterplot for the “greater than” signature (e.g., “>”) such that variance in the rate is larger at small population sizes (Fig. 67.1). Finally, brush select on the scatterplot to see where the areas with high rates and low population sizes appear on the map. These are the places with apparent high rates that may be unstable due to the small numbers problem.

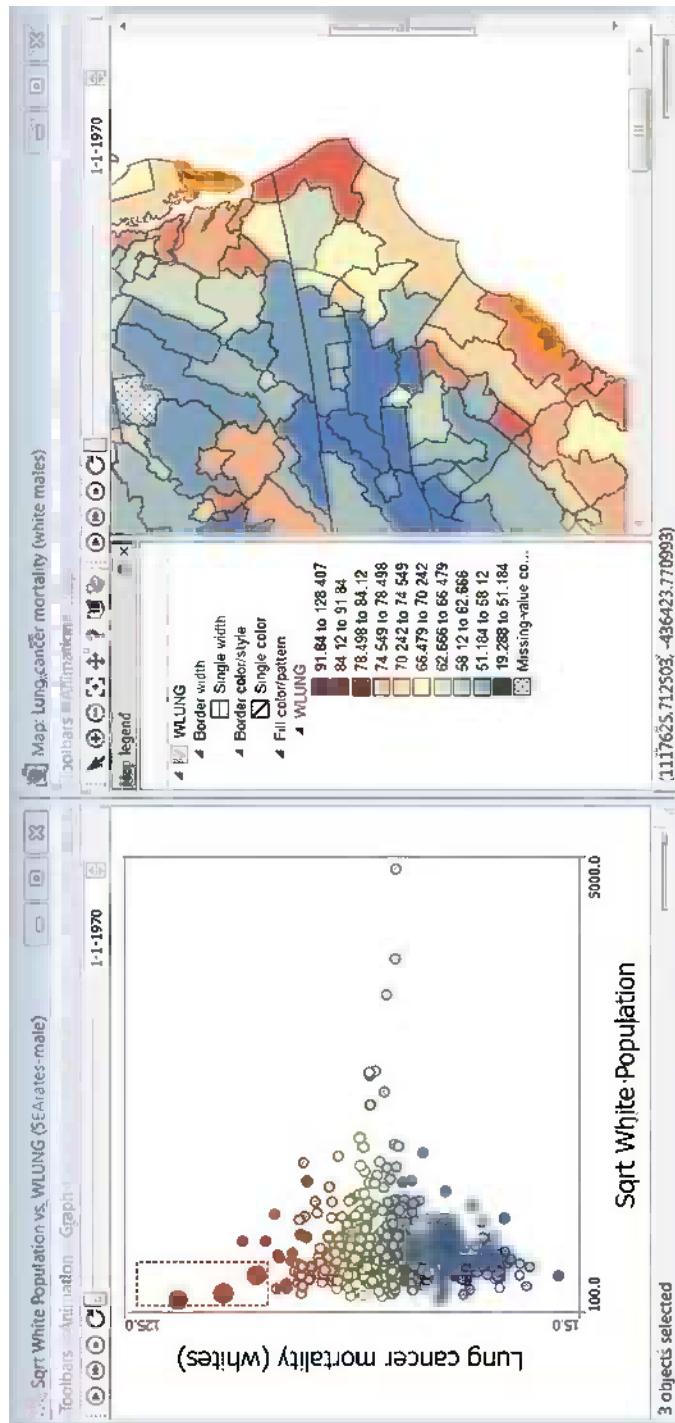


Fig. 67.1 Simple diagnostic for the small numbers problem. A plot of the lung cancer mortality rate for white males (y-axis) versus the square root of the white male population (x-axis) demonstrates the “>” signature, with higher variance in the rate at small population sizes. Brush selection on the scatterplot (the three large red circles in *dashed rectangular box*) locates the areas with high mortality rates that may be unstable since they have small denominators. Calculated in BioMedware SpaceStat software

Variability in rates due to the small numbers problem, if not corrected for, can give rise to artifactual spatial structure in the estimated rates. For example, the three areas with high rates brush selected in Fig. 67.1 are high spatial outliers. When clustering rates, it therefore is important to use statistical techniques that either stabilize the rates by constructing local populations with similar denominator sizes or that account for denominator size when assessing statistical significance.

Constructing Local Populations with Similar Denominator Sizes. A classic example of the former approach is Turnbull's test, which uses spatially adaptive kernels to construct local populations with common denominator sizes (Turnbull et al. 1990). This method scans populations within the study area for clusters of cases. A circular window is centered on each region in turn and expanded to include neighboring regions until the total aggregated population within the window equals a user-defined threshold, R . One can think of this as "borrowing strength" from neighboring areas to construct local populations comprised of the same at-risk population size. These circular windows may overlap, and the counts within the windows will not be independent. The test statistic, M_R , is the maximum number of cases observed among all windows of population size R .

Accounting for Denominator Size When Assessing Statistical Significance. A widely used example of the latter approach is the heterogeneous Poisson model, which is used to specify null spatial models for inferential statistics and as the point of departure for more complex modeling approaches. A Poisson disease process is described by a parameter, lambda (λ), often referred to as the intensity, and the counts of the health events in local areas are assumed to be samples from a Poisson distribution with a given intensity and population size in each local area:

$$y_i \sim \text{Poisson}(\lambda_i, E_i) \quad i = 1, 2, \dots, A \quad (67.1)$$

Here, y_i denotes the observed count of health events (e.g., incident cases or deaths) in area i , A is the number of areas (e.g., counties or census tracts) under consideration, and E_i is the expected count of health events based on the size of the population at risk in area i after correction for known risk factors and covariates (e.g., age structure). λ_i , then, is an area-specific relative risk variable, defining the relative disease risk in the population residing in area i . This model is *heterogeneous* since the disease risk (the λ_i) may vary from one area to another.

A null hypothesis frequently used when evaluating counts of clusters of health events in local areas is that the underlying disease risk is the same from one area to another so that area-specific elevations in risk are absent. Here the null hypothesis could then be written:

$$H_0 : \lambda_1 = \lambda_2 = \dots = \lambda_n; \quad i = 1, \dots, A \quad (67.2)$$

Global, Focused, and Local Tests for Clusters of Health Events. We then might proceed by evaluating the "omnibus" alternative hypothesis of rejecting the null hypothesis (here " $-$ " is negation):

$$H_1 : \neg(H_0). \quad (67.3)$$

In health event clustering, this corresponds to what is known as a global test for clustering such that rates are elevated somewhere in the study area, but the null hypothesis does not specify which areas are elevated. Focused cluster tests use more detailed alternative hypotheses that state that certain areas might have elevated risk. These might be chosen, for example, to be near a known source of carcinogens. Local tests search for elevated risk in specific areas, but do not require prior knowledge or specification of which areas might have elevated risk.

Interpolation Autocorrelation. Smoothing rates in an attempt to adjust for rate instability can introduce spatial autocorrelation due to interpolation. Smoothing introduces nuisance autocorrelation whenever the kernels or models used to accomplish the smoothing overlap. Examples include inverse distance and empirical Bayesian interpolation. Here, the spatial scale of the autocorrelation introduced by smoothing will depend on the kernel size. When assessing clusters, it may be inappropriate to cluster rates after first smoothing them, since the smoothing step can introduce artifactual local similarity in rates attributable to interpolation rather than to underlying disease processes. One thus may wish to use smoothing when displaying maps of the rates, but employ techniques that explicitly account for denominator size when evaluating clustering.

Access to Screening, Care, and Treatment. Access to health-care and screening facilities can give rise to spatial autocorrelation in health events since both screening and treatment influence health outcomes. For example, several studies have demonstrated that access to breast cancer screening facilities is significantly associated with geographic differences in stage at diagnosis, with late-stage cancers more frequent in populations distant from breast cancer screening facilities. Poorer populations are particularly impacted by access to screening, since availability of transport and travel times may pose barriers to seeking health screening. An example is the use of mosquito nets, malaria incidence, and distance to clinics that distribute the nets (Enayati and Hemingway 2010). In agrarian rural areas of Malawai with poor roads, a distance of 10 km to the nearest clinic where mosquito nets are distributed may involve a full day round trip. Not surprisingly, studies have demonstrated that households nearer to clinics have higher mosquito net usage rates than households that are distant. A useful intervention then is to distribute the mosquito tents directly to the households.

Neighborhood/Contextual Effects. Neighborhood and related contextual effects can have negative impacts on human health status that exceed the impacts of covariates such as socioeconomic status and access to care that themselves may vary dramatically from one neighborhood to another (Spielman and Yoo 2009). Hypotheses suggest that perception of personal safety and quality of the neighborhood living environment can result in chronic stress that leads to reduced immune function and increased disease susceptibility, elevated blood pressure, and heart disease. One mechanism is the interaction between chronic stress, elevated cortisol, and immune system status, such that chronic stressors are associated with

suppression of both cellular and humoral measures of immune system function. Neighborhoods thus may be associated with spatial autocorrelation in health effects through direct effects such as socioeconomic determinants (e.g., income and health insurance), environmental factors (such as air quality), as well as contextual effects that impact stress and immune function.

Differences in Response to Health-Care Policy. Policies related to health care, treatment, drug development and deployment, and care delivery can have substantial impacts on health outcomes that may differ from one geographic area to another. In the United States, the states have a fair amount of flexibility in how they implement national policies. For example, the Center for Disease Control (CDC) is required to conduct the Behavioral Risk Factor Surveillance System (BRFSS), which is an ongoing telephone health survey system, tracking health conditions, and risk behaviors in the United States annually since 1984. Data are collected monthly by all 50 states, the District of Columbia, Puerto Rico, the U.S. Virgin Islands, and Guam. A core portion of the health survey questions come from the CDC, but states can supplement the survey with their own optional modules, and the BRFSS variables may thus vary from one state to another.

In addition, health policies can have differential impacts on physician behaviors that are not immediately apparent when the policies are drafted. For example, a recent study explored geographic variation in the use of physician-administered chemotherapeutic agents under Medicare Part B in response to a major reform of Medicare's reimbursement system (Jacobson et al. 2011). Physician prescription behavior in response to the payment change varied from state to state. Some states increased treatments with certain chemotherapeutic agents by 4 %, and a few actually reduced treatment rates. The state-to-state differences are statistically significant, with the null hypothesis that the change in chemotherapy treatment was the same across states rejected at $p < 0.001$ level.

Healthy Worker and Geographic Attractors. The "healthy worker effect" describes the reduced disease risk observed among employed individuals in many industries that cuts across different diseases. This can give the false appearance of no differences in risk between workers employed in a given industry when compared to the larger population, even though substantial occupational risks may be present (Fornalski and Dobrzański 2010). Workers tend to follow employment opportunities, and the establishment of large manufacturing facilities can attract a cohort of healthy workers, resulting in an apparent deficit of disease risk in neighborhoods where these workers reside. A related phenomenon is that of the "geographic attractor" that arises after health conditions are diagnosed. Here, individuals decide to move nearer hospitals, clinics, and treatment centers to ease health-care access. When they die, the place of death is recorded as their last known residence, leading to an apparent excess of disease near treatment facilities.

Outbreaks/Spread of Infection. Infectious diseases transmitted through the air, sexual contact, fomite transmission, by drinking water contaminated with pathogens, and other means often require infected and susceptible individuals to be in close proximity to one another. This is true for pathogens with limited life spans outside the human body (e.g., influenza viruses), but is less true for those with

a dormant phase that can survive outside the body for extended periods (such as anthrax spores). For highly infectious pathogens transmitted from person to person, we may observe an initial outbreak from an index case (the first case to appear in a local population) that is followed by a spatial “wave” of infection that moves outward from the location of the index case. This may be followed by an endemic phase characterized by the maintenance of lower levels of infection in the population characterized by local outbreaks or by the infection spreading rapidly and dying out. Geographic pattern in the spread of infection is mediated by complex interactions between the probability of infection transmission, the contacts between infected and susceptible individuals, the life history of infection including the duration and timing of the infective stage, mobility of infected and susceptible individuals, timing of the rise and waning of immunity, the virulence of infection, as well as other factors (Sattenspiel and Lloyd 2010).

Immunity. When considering spatial autocorrelation in the spread of infectious diseases, the geography of immunity can be an important consideration. Issues include the waning of immunity, herd immunity, vaccination behaviors, and vaccine availability and distribution (Funk et al. 2010). When pathogens enter the body, the immune system develops antibodies to fight the infection. Immune response is said to wane as the concentrations of antibodies specific to that pathogen decrease over time. When immunity has waned sufficiently, the person may then become infected once again. This process can result in the appearance of clusters where members of a local population are infected, become immune, and then a resurgence of infection as immunity wanes, resulting in space-time patterns in infection.

Vaccination confers immunity without having to undergo a full-blown infection. *Herd immunity* is the protection from infection that arises when a sufficiently large proportion of the population has been vaccinated. Infection transmission halts when enough individuals are vaccinated and are immune, conferring protection even to those who have not been vaccinated. Vaccination itself often follows geographic distribution and adoption patterns. Hence, the vaccine distribution strategy can impact the timing of when immunity is conferred by vaccination and thus the geographic spread of infection. Vaccination itself can have intriguing side effects in terms of disease ecology. The eradication of small pox is one of the great public health triumphs of our time, in which the global distribution and administration of the smallpox vaccine eradicated the disease in human populations. Immunity to smallpox confers partial immunity to a related infection, monkey pox. Once smallpox was eradicated, the smallpox vaccination program stopped, and outbreaks of monkey pox infections are now increasing (Rimoin et al. 2010).

Geographic Variation in Positional Error. That errors in case ascertainment and incomplete reporting can complicate the detection of disease clusters is well known (Kingsley et al. 2007). Positional error can also impact cluster detection, in at least two ways. First, geographic confounding arises when geographic variation in risk factors is associated with geographic variability in positional error. The potential for this is larger than one might expect as positional error in geocoded place of residences is larger in rural areas, a gradient similar to certain environmental risk factors and socioeconomic and demographic variables. Second, positional error decreases the

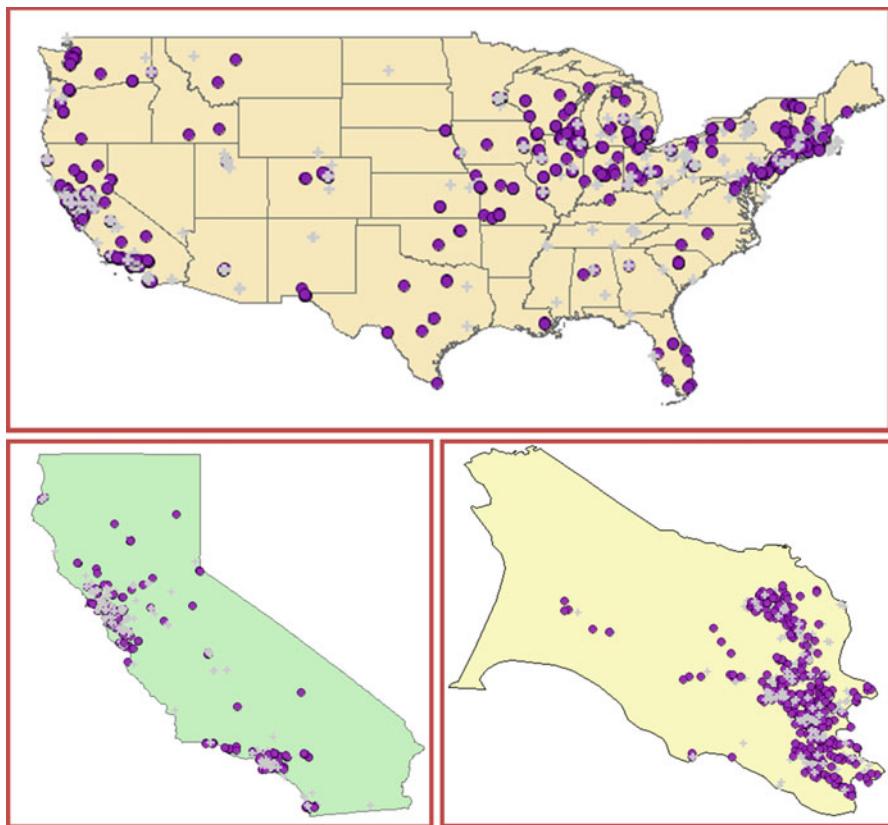


Fig. 67.2 Locations of places of residence of breast cancer cases (circles) and controls (plus symbols). Geographic locations of place of residence may vary dramatically from that observed at time of diagnosis in Marin county (*lower right*) to where women lived over their life course in the US (*top*) and California (*lower left*) (Source: Jacquez et al. 2011)

power to detect true clusters. Hence, our ability to detect clusters will vary geographically when gradients in positional errors are present.

Migration/Latency. Both chronic and infectious diseases have a latency between causative exposures and diagnosis. For cancers, this latency can be a decade or more, for infectious diseases such as influenza, it may be days. Because humans are mobile, the geographic pattern of where individuals were when they were exposed may differ dramatically from where they were when they were diagnosed. Consider the example of breast cancer. Breast cancer is a complex disease thought to have long latencies on the order of decades, although a small proportion of cases do appear in childhood and adolescence. The geographic pattern of where women lived over their life course differs dramatically from where they lived when they were diagnosed (Fig. 67.2). For many health outcomes, geographic patterns in cases at time of diagnosis may differ dramatically from that observed at disease onset.

67.6 Role of Neutral Models

The above paragraphs provided an overview of some of the sources of spatial autocorrelation in health events. When exploring disease patterns and clusters, many of these sources of geographic variation may not be of direct interest, for example, we often may wish to account for spatial heterogeneity in population density when searching for the signature of causative exposures underlying clusters of disease. Here, the idea is to search for clusters of health events above and beyond that attributable to geographic variation in population density. This concept can apply to any source of geographic pattern that may not be of direct interest; those members, for example, of the set of possible explanations that are described earlier under “Strong inference” and enumerated in [Sect. 67.5](#).

When clustering health events, one then incorporates geographic variability in covariates and other factors not considered to be of interest into the null hypothesis. Mechanistically, this usually is accomplished using approximate randomization that includes observed variation patterns in those factors not of direct interest into the null spatial model. Models that accomplish this have been referred to as “neutral” rather than “null” models, to capture the idea that they account for more than just “complete spatial randomness.” Neutral models thus correspond to plausible system states that can be used as a reasonable null hypothesis (e.g., “background variation”) in disease cluster tests. The problem then is to identify spatial patterns above and beyond that incorporated into the neutral model, enabling, for example, the detection of clusters *above and beyond* background or regional variation in the risk of developing disease.

A typology of neutral models that account for factors often encountered in analyses of health events defines neutral models type I–VI. These neutral models are realistic in that they account for the spatial autocorrelation, nonuniform risk, and spatially heterogeneous population sizes that may be present in the absence of the cluster process. Model I is complete spatial randomness (CSR) that is still widely used in health analysis even though it usually does not correspond to any plausible state of the system being studied. Model II reproduces the spatial autocorrelation that may be present in the observed data. Model III incorporates nonuniform variability in the underlying risk that may be attributable to risk factors and covariates that are not of direct interest. Models IV to VI account for the impact of population size and variability on the stability of observed rates and are used to address the small numbers problem.

Neutral models thus play a critical role in scientific inference in disease pattern analysis since they allow one to systematically incorporate different sources of geographic variation, including spatial autocorrelation, into the hypotheses being evaluated. In the framework of Strong inference, one conducts a series of statistical analyses systematically evaluating each of the hypotheses in the set of alternative explanations for the observed spatial patterns of health events. These are each incorporated into the neutral model of a given spatial cluster test, and if the test is significant, that hypothesis is rejected; if it is not significant, that hypothesis is retained in the set of plausible explanations for the observed spatial pattern.

An alternative mechanism when knowledge of the system is sufficient is to construct more formal, detailed models using (spatial) regression with in a maximum likelihood or Bayesian framework, or other modeling approaches. The variability captured by the model is then attributable to the predictor variables, and clustering may then be applied to the regression residuals to quantify spatial pattern not captured by the model itself.

67.7 Data: What to Cluster?

Several different types of health events are commonly analyzed, including incident cases and deaths, diagnoses and stage at diagnosis of cancers and other conditions, symptoms, volunteered or crowd-sourced data, Web search terms, pathogen strains, mobility traces, and model residuals.

Incident cases may come from disease reporting systems such as cancer, birth defect, and infectious disease registries. This information may include date of diagnosis, place of residence at time of diagnosis, place of diagnosis, and patient identifiers. The patient identifiers and place of residence can be very useful in a spatial analysis, as they support linking of the diagnostic information to other data on Medicare usage, socioeconomic status, and behavioral risk factors, among others. This data linking may be accomplished by the registry, in order to protect access to data and confidentiality, or it may be accomplished by the researchers. Geographic bias may be introduced when case reporting varies from one geographic area to another, in which case areas with more complete reporting may appear as aberrant spatial clusters due to sample bias.

Incident death data often is of higher quality and can be more complete than case data, as all deaths must be reported to state and public health agencies. These typically include cause of death, identifiers for the deceased, and place of residence at time of death. As noted earlier, competing causes of death may result in underreporting of comorbid conditions, as when a person dies of lung cancer but also has Alzheimer's. Here, the cause of death may be attributed to lung cancer. Both incidence and mortality data may lag one or more years between when the health events occurred and when the data are available from the registry/reporting agency. This allows for case validation, data entry, and checking for completeness.

Diagnoses and stage at diagnosis are used to conduct surveillance, to identify populations with advanced stages of disease, and to aid in the siting of screening facilities. Diagnoses for infectious diseases of public health interest such as influenza, HIV, sexually transmitted diseases, and others may be reported to local and state health agencies and from there to centralized reporting facilities such as the Centers for Disease Control. This information may be timely for infectious diseases, allowing weekly assessment of outbreaks and infection trends. Cancer staging information may be available describing tumor status from localized (*in situ*) to metastasized/advanced and provide important advantages for spatial analyses. First, they may be used to assess disparities in stage at diagnosis, which

can provide insights into differential access to screening facilities and/or differences in exposures and genetic predispositions. Second, they also have the advantage of providing ready-to-use denominators, since the population at risk will be the total number of diagnoses summed across stages.

Symptoms may be quite useful for undertaking syndromic surveillance that seeks to identify the early onset of infectious diseases, outbreaks, and bioterrorism attacks. Here, the incoming data may be of uncertain quality, including over-the-counter sales at pharmacies for symptom-related products such as antihistamines, cold medicines, and diarrhea medications. More recently, syndromic surveillance approaches have been applied to data from Google Web search terms and to crowd-sourced and volunteered information on disease outbreaks, with surprising accuracy in terms of the predictive ability, even though a practical means for validating the incoming data is frequently absent (Adams 2011). This is expected to be a growth area in applied spatial analysis in public health, as it can provide near real-time data streams at nominal expense.

The genetic similarity of *pathogen strains* can be closely associated with how far apart the infected individuals are on the chain of infection, due to the population bottlenecks that occur with each transmission event and to the increasing number of replications of the pathogen as distance on the chain of infection increases. Replications increase genetic change as replication errors and genetic mutations accumulate. This can lead to strong spatial autocorrelation in within-patient pathogen strain genetics, such that nearby infected individuals tend to be infected with similar pathogen strains. Polymerase Chain Reaction and other technologies for rapid molecular genotyping of pathogen strains are now being used for hospital infection surveillance and outbreak detection (Ecker et al. 2009). The emergence of antibiotic- and drug-resistant pathogen strains is thought attributable to strong selective pressures posed by antibiotic treatment, and distinct multidrug-resistant pathogen strains are known to emerge in specific clinics and geographic settings. Hence, spatial patterns in pathogen genetics can provide useful information regarding pathogen emergence, history, and infection transmission. Coevolutionary theory for pathogens and their hosts results in distinct predictions regarding how pathogen and host genetics may change through time in different places, although complex pathogen life histories and other microevolutionary processes can obscure the expected similarities.

Spatial autocorrelation in the *residuals* from spatial models treating health events as dependent variables and various risk factors and covariates as independent variables (risk models) is often used as a model diagnostic. Spatially autocorrelated residuals violate assumptions of regression (iid, independent and identically distributed residual errors) and may indicate a missing predictor that is itself spatially structured or a spatial process, such as migration, that is not present in the model. For risk models, the residual is the difference between the observed risk (e.g., incidence and mortality rate) and the modeled risk. When the residual is positive, the model underpredicts the observed risk; hence, local populations with positive residuals have excess risk beyond that explained by the model.

67.8 Data Models and Clustering Methods

There are several reviews of cluster and cluster modeling of health events including inferential statistics, geostatistics, regression, and Bayesian techniques (Goovaerts 2009; Lawson and Banerjee 2009; Rogerson 2006; Waller and Gotway 2004; Aldstad 2010). Here we provide a quick overview of several data models and corresponding clustering methods. The reader may wish to visit the above references and other chapters in this handbook for further details on clustering methods.

Individual health events, such as the place of residence and time of diagnosis for a case of leukemia, provide the basic datum of the form:

$$x_i, y_i, t_i \quad (67.4)$$

This represents health events as points in space-time and assumes events have no duration and may be assigned to discrete locations. Each record usually corresponds to a disease case, and there are n total health events (cases). This data model has been criticized as overly simplistic, since it ignores disease latency and assumes individuals do not move from place to place.

This data model underlies many tests for space-time interaction in health events. Interaction tests search for patterns such that nearby health events occur at about the same time, a pattern taken to be indicative of contagious processes (e.g., infection) and local elevations in disease risk that occur over discrete times. Examples of tests for interaction include the Mantel, Knox, and k-NN tests. Here, we give a quick definition of Mantel's test.

The Mantel test calculates s_{ij} and t_{ij} as the spatial and temporal distances, respectively, between two health events i and j . These distances may be scaled to reduce the impact of larger distances on the test statistic. Mantel's statistic is the sum of the products of the space and time distances:

$$Z = \sum_{i=1}^n \sum_{j=1}^n s_{ij}t_{ij} \quad (67.5)$$

This often is scaled to yield a correlation:

$$r = \frac{1}{(n^2 - n - 1)} \sum_{i=1}^n \sum_{j=1}^n \frac{s_{ij} - \bar{s}_s}{s_s} \frac{t_{ij} - \bar{t}_t}{s_t} \quad (67.6)$$

Here, \bar{s} is the average space distance, \bar{t} is the average time distance, and s_s and s_t are the standard deviations of the spatial and time distances, respectively. The standardized measure is a matrix correlation that is in the range -1 to 1 , with large positive values indicating significant interaction such that nearby health events occur at about the same time and with negative values indicating avoidance, so that nearby health events tend to occur in temporal isolation.

Counts, for example, of the number of incident cases in each area, arise when denominator information is lacking, preventing the counts from being converted

into rates. Here, the underlying data model may be written simply as f_{it} , which is the count of health events in area i over a defined temporal support (e.g., count of incident cases of leukemia in county i in 2010). Since areas will differ in size of the at-risk population, spatial analysis of counts across areas will not be particularly informative at a single point in time. Count data through time are useful for surveillance when one is interested in assessing whether to sound a “cluster alarm” when the counts increase relative to a historic baseline (Höhle and Paul 2008), assuming the size of the underlying at-risk population is relatively static. Rates are comprised of a numerator and denominator with the count of the number of health events in the numerator and the population at risk in the denominator. Examples include incidence, mortality, morbidity, and accident rates. Raw rates may be of little use in spatial analyses since they do not correct for important covariates, such as age. It is not informative, for example, to search for clusters of childhood leukemia in retirement communities where only those 55 and older may reside. Rates therefore are often age-standardized, and standardization may be accomplished for many of the covariates identified earlier (Greenland 2004). There are many techniques for analyzing clustering of data defined by numerators and denominators, and these often employ a heterogeneous Poisson model with uniform risk as the underlying null hypothesis, as described earlier. Among the most widely used are spatial scan statistics that may use circular-, elliptical-, and flexibly shaped scanning kernels (Kulldorff et al. 2007; Tango and Takahashi 2005). Boundary analysis approaches that work at higher spatial resolution in that they are sensitive to changes across the area edges are also available (Lu and Carlin 2005). Spatially explicit *health disparity* statistics may be used when numerator and denominator data are available for two groups, one a target (such as black males) and the other a reference population (such as white males). These statistics seek to identify statistically significant differences in local population means for the two groups. Examples include the analysis of racial disparities in lung cancer mortality in US Congressional districts (Gallagher et al. 2009).

Data from *case-control* studies are often modeled as marked spatial point processes with two states: case and control. One then can write a data model as

$$x_i, y_i, c_i \quad (67.7)$$

Here, x_i , y_i are the geographic coordinates of the health event for study participant i , and c_i is a case-control identifier defined as

$$c_i = \begin{cases} 1 & \text{if participant } i \text{ is a case} \\ 0 & \text{if participant } i \text{ is a control} \end{cases} \quad (67.8)$$

Several techniques have been proposed for cluster analysis and surveillance analysis with case-control data; these all account for geographic heterogeneity in population density. Cuzick and Edwards test (1990) is based on nearest neighbor relationships:

$$T_k = \sum_{i=1}^n \sum_{j=1}^k c_i c_j \quad (67.9)$$

Here, T_k is the test statistic, which is large when cases tend to be nearest neighbors of other cases; k is the number of nearest neighbors being considered, analogous to a kernel size but defined by the local number of nearest neighbors to evaluate (e.g., ten nearest participants to case i); and the inner summation iterates over the k nearest neighbors of participant i .

The cases and controls are assumed to be sampled from the same underlying spatial distribution; hence, there should not be clustering of the cases relative to the controls. If there is, this may be the signature of an underlying spatial process such as infection/contagion, localized exposures to factors that increase disease risk, case attractors (such as clinics as mentioned earlier), and geographic sampling bias. When analyzing data from case-control studies, it is important that the sampling strategy *not* be geographically stratified, as this will tend to obscure true clustering of cases relative to controls. To see this, suppose the controls are sampled proportionately to the cases in subparts of the study area, such as townships. For every case that occurs in a township, the sample is geographically stratified to draw a control from that township. This makes it impossible to detect a geographic clustering at the township scale. The data model is static and ignores mobility of the study participants, a need that is addressed using mobility histories.

Mobility histories model traces of location-enabled devices, such as smart phones, as well as residential histories of study subjects and industries. For example, residential histories can be represented as space-time step functions ([Fig. 67.3](#)). To identify the location and timing of significant clustering of cases relative to controls using this data model, a spatially and temporally local case-control cluster statistic may be defined as

$$Q_{i,t}^{(k)} = c_i \sum_{j=1}^k \eta_{i,j,t}^{(k)} c_j \quad (67.10)$$

This quantity is the count, at time t , of the number of k nearest neighbors of case i that are cases, and not controls. Individuals i and j have case-control identifiers, c_i and c_j , as defined earlier. The term $\eta_{i,j,t}^{(k)}$ is a binary spatial proximity metric that is one when participant j is a k nearest neighbor at time t of participant i ; otherwise, it is zero. Since a given individual i may have k unique nearest neighbors, the $Q_{i,t}^{(k)}$ statistic is in the range $[0, k]$. When i is a control, $Q_{i,t}^{(k)} = 0$. When i is a case, low values indicate cluster avoidance (e.g., a case surrounded by controls), and large values indicate a cluster of cases. When $Q_{i,t}^{(k)} = k$, at time t all of the k nearest neighbors of case i are cases.

With this local statistic defined, several global, local, and focused cluster tests that account for residential mobility can be defined by integrating over a person's life course, over space (a time slice), and over time. When integration is accomplished

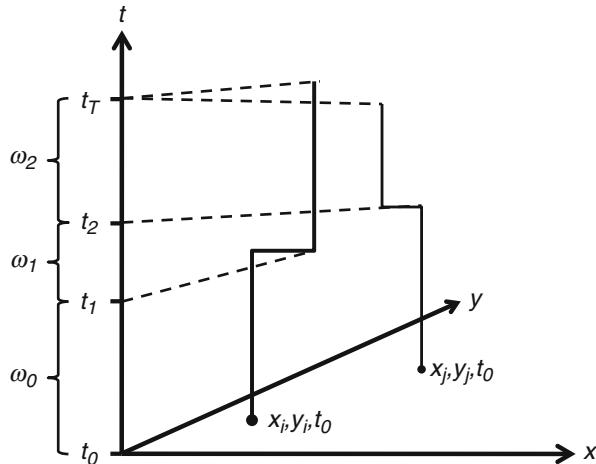


Fig. 67.3 Residential histories as space-time step functions. The axes x and y define a geographic domain (e.g., longitude and latitude decimal degrees), the t axis represents time (e.g., date). The study extends from time t_0 to time t_T . The residential histories for persons i and j are shown as step functions through space-time. For example, person i begins the study residing at location x_i, y_i, t_0 . They remain at that geographic coordinate until the instant before time t_1 , when they move to x_i, y_i, t_1 . The duration of time they reside at this first place of residence is ω_0 .

over a subject's residential history, we think of this as a "life course" statistic that assesses a tendency to have other cases rather than controls nearby through time.

$$Q_i^{(k)} = \int_{t=t_0}^T Q_{i,t}^{(k)} dt \quad (67.11)$$

A time-specific statistic that provides an overall measure of case clustering when all of the participants are considered together is given in Eq. (67.12). It is the sum, over all cases, of the subject-specific and time-specific measure of case clustering in Eq. (67.10).

$$Q_t^{(k)} = \sum_{i=1}^{n_1} Q_{i,t}^{(k)} \quad (67.12)$$

Analogous to Cuzick and Edwards test Eq. (67.9), $Q_t^{(k)}$ evaluates global clustering of cases at time t , such that the amount of case clustering observed when all of the participants are considered together is evaluated. For convenience, the summation is over the n_1 cases. This is a "global" test since it is comprised as the sum of the local statistic from Eq. (67.10).

$$Q^{(k)} = \sum_{i=1}^{n_1} Q_i^{(k)} \quad (67.13)$$

Equation (67.13) provides a global test that sums over all of the life course statistics from Eq. (67.11). It is used to evaluate whether there is significant life course clustering of cases when all of the participants are considered together.

Mobility histories have the advantage of relaxing the assumption of immobile individuals imposed by methods based on static spatial point distributions. Residential histories for health studies are available in the USA from commercial vendors and appear to have acceptable accuracy (Jacquez et al. 2011). Several European countries, such as Denmark, track place of residence information from birth, a valuable resource in spatial health analysis.

67.9 Conclusions

This chapter has focused on spatial autocorrelation and clustering of health events. It presented a cognitive framework for the spatial analysis of health events based on Strong inference, which systematically evaluates the hypotheses that might explain an observed spatial pattern. Operationally, each hypothesis may be incorporated into the null hypothesis of a statistical test for disease pattern using neutral models, and in this manner, the members of the set of plausible hypotheses can be systematically evaluated. The remaining hypothesis(es) then plausibly explains the observed geographic pattern in health events.

Sources of positional uncertainty need to be carefully evaluated when assessing disease clusters, as they are now known in some instances to be substantial, potentially leading to false negatives. Under Strong inference, this can lead one to incorrectly accept the neutral model, thereby retaining it as an explanation for the observed spatial patterns when it should, in fact, be excluded.

Philosophically, the spatial analysis of health events makes an implicit assumption regarding the importance of location for the health event in question. As noted earlier, place of residence at time of diagnosis can be misleading when disease latency is long, especially when one wishes to make inferences regarding disease onset and causative exposures. Aside from latency processes, there are other reasons why location may not be too informative, such as whether it is appropriate to use location as an exposure surrogate. Spatial analysis contrasts in some respects with the more traditional clinical approach that focuses on the individual as the nexus of health data. Ultimately, we hope to make inferences regarding the causes of disease and changes in health status at the individual level and to be able to extend those findings to put in place appropriate public health policies.

The research funding, technological, and data milieu are changing rapidly. In an era of constrained funding, large case-control, cohort, and other study designs will be increasingly difficult to undertake. At the same time, new data sources are becoming available, and the era of “Big data” is upon us, where streams from Web searches,

smart phones, houses with Web-enabled devices, and a broad range of location- and Web-enabled devices are fused and mined. Big data hold the possibility of replacing the expensive large epidemiological study with studies incorporating crowd sourcing, volunteered geographic information, and other sources. But this will require advances in spatial epidemiological study designs that put in place appropriate sampling frameworks for Big data that support sound scientific inference. This is one of the new frontiers at the forefront of spatial health analysis.

Acknowledgments The author's efforts were funded in part by grants 2R44CA112743, 5R44CA135818, and 1R21LM011132 from the National Cancer Institute and the National Library of Medicine. The perspectives are those of the author and do not necessarily represent those of the funding agencies.

References

- Adams SA (2011) Sourcing the crowd for health services improvement: the reflexive patient and "share-your-experience" websites. *Soc Sci Med* 72(7):1069–1076
- Aldstad J (2010) Spatial clustering. In: Fischer MM, Getis A (eds) *Handbook of applied spatial analysis. Software tools, methods and applications*. Springer, Berlin/Heidelberg/New York, pp 270–300
- Cuzick J, Edwards R (1990) Spatial clustering for inhomogeneous populations. *J R Stat Soc B* 52(1):73–104
- Ecker DJ, Massire C, Blyn LB, Hofstadler SA, Hannis JC, Eshoo MW, Hall TA, Sampath R (2009) Molecular genotyping of microbes by multilocus PCR and mass spectrometry: a new tool for hospital infection control and public health surveillance. In: *Molecular epidemiology of microorganisms*. Springer, Berlin/Heidelberg/New York, pp 71–87
- Enayati A, Hemingway J (2010) Malaria management: past, present, and future. *Annu Rev Entomol* 55(1):569–591. doi:10.1146/annurev-ento-112408-085423
- Fornalski KW, Dobrzański L (2010) The healthy worker effect and nuclear industry workers. *Dose-Response* 8(2):125–147
- Funk S, Salathé M, Vincent A, Jansen A (2010) Modelling the influence of human behaviour on the spread of infectious diseases: a review. *J R Soc Interface* 7(50):1247–1256. doi:10.1098/rsif.2010.0142
- Gallagher CM, Goovaerts P, Jacquez GM, Hao Y, Jemal A, Meliker JR (2009) Racial disparities in lung cancer mortality in U.S. congressional districts, 1990–2001. *Spat Spattemporal Epidemiol* 1(1):41–47. doi:10.1016/j.sste.2009.07.007
- Getis A (2010) Spatial autocorrelation. In: Fischer MM, Getis A (eds) *Handbook of applied spatial analysis. Software tools, methods and applications*. Springer, Berlin/Heidelberg/New York, pp 255–278
- Goodchild MF, Alan Glennona J (2010) Crowdsourcing geographic information for disaster response: a research frontier. *Int J Digit Earth* 3(3):231–241
- Goovaerts P (2009) Medical geography: a promising field of application for geostatistics. *Math Geol* 41(3):243–264
- Greenland S (2004) Model-based estimation of relative risks and other epidemiologic measures in studies of common outcomes and in case-control studies. *Am J Epidemiol* 160(4):301–305. doi:10.1093/aje/kwh221
- Höhle M, Paul M (2008) Count data regression charts for the monitoring of surveillance time series. *Comput Stat Data Anal* 52(9):4357–4368. doi:10.1016/j.csda.2008.02.015

- Jacobson M, Earle CC, Newhouse JP (2011) Geographic variation in physicians' responses to a reimbursement change. *N Eng J Med* 365(22):2049–2052. doi:10.1056/NEJMp1110117
- Jacquez GM, Slotnick MJ, Meliker JR, AvRuskin G, Copeland G, Nriagu J (2011) Accuracy of commercially available residential histories for epidemiologic studies. *Am J Epidemiol* 173(2):236–243
- Johnson G, Buckeridge D, Dearth S, Ditty J, Finelli L, Hopkins RS, Hummel J, et al. (2012) Draft guidelines for syndromic surveillance using inpatient and ambulatory clinical care EHR data. A report from the international society for disease surveillance: international society for disease surveillance
- Kingsley BS, Schmeichel KL, Rubin CH (2007) An update on cancer cluster activities at the centers for disease control and prevention. *Environ Health Perspect* 115(1):165–171
- Kulldorff M, Mostashari F, Duczmal L, Katherine Yih W, Kleinman K, Platt R (2007) Multivariate scan statistics for disease surveillance. *Stat Med* 26(8):1824–1833
- Lawson AB, Banerjee S (2009) Bayesian spatial analysis. In: Fotheringham S, Rogerson P (eds) *The handbook of spatial analysis*. Sage, London, pp 321–342
- Lu H, Carlin BP (2005) Bayesian areal wombling for geographical boundary analysis. *Geogr Anal* 37(3):265–285
- Neutra RR (1990) Counterpoint from a cluster buster. *Am J Epidemiol* 132(1):1–8
- Platt JR (1964) Strong inference. *Science* 146:347–353
- Rimoin AW, Mulembakani PM, Johnston SC, Lloyd JO, Smith NK, Kisalu TL, Kinkela SB et al (2010) Major increase in human monkeypox incidence 30 years after smallpox vaccination campaigns cease in the Democratic Republic of Congo. *Proc Natl Acad Sci* 107(37):16262–16267. doi:10.1073/pnas.1005769107
- Rogerson PA (2006) Statistical methods for the detection of spatial clustering in case-control data. *Stat Med* 25(5):811–823
- Sattenspiel L, Lloyd A (2010) *The geographic spread of infectious diseases: models and applications*. Princeton University Press, Princeton
- Spielman SE, Yoo E (2009) The spatial dimensions of neighborhood effects. *Soc Sci Med* 68(6):1098–1105. doi:10.1016/j.socscimed.2008.12.048
- Tango T, Takahashi K (2005) A flexibly shaped spatial scan statistic for detecting clusters. *Int J Health Geogr* 7(14):11
- Turnbull BW, Iwano EJ, Burnett WS, Howe HL, Clark LC (1990) Monitoring for clusters of disease: application to leukemia incidence in upstate New York. *Am J Epidemiol* 132(1 Suppl): S136–S143
- Waller LA, Gotway CA (2004) *Applied spatial statistics for public health data*. Wiley, Hoboken
- Weissman JS, Hasnain-Wynia R (2011) Advancing health care equity through improved data collection. *N Eng J Med* 364(24):2276–2277

Mariana Arcaya and S. V. Subramanian

Contents

68.1	Introduction	1336
68.2	Recognizing Multilevel Data Structures	1337
68.3	Causal Research Questions That Motivate Multilevel Analyses	1339
68.3.1	Specific Ecologic Effects	1339
68.3.2	Common Ecologic Effects	1340
68.3.3	Composition Versus Context	1341
68.3.4	Describing Contextual Heterogeneity	1342
68.3.5	Characterizing and Explaining Contextual Variations	1343
68.4	Multilevel Framework: A Necessity for Understanding Ecologic Effects	1343
68.5	Multilevel Statistical Analysis	1345
68.5.1	Variance Component or Random Intercepts Model	1346
68.5.2	Contrasting Multilevel and Other Approaches: Fixed Versus Random Effects	1348
68.5.3	Contrasting Multilevel and Other Approaches: Marginal Versus Mixed Approaches	1353
68.6	Sources of Bias in Ecologic Inference	1354
68.6.1	Reverse Causation	1354
68.6.2	Uncontrolled Confounding	1355
68.6.3	Covariate Balance	1355
68.6.4	Treatment Heterogeneity	1355
68.6.5	Alternatives to Observational Data	1356
68.7	Extensions to the Basic Multilevel Models	1357
68.7.1	Multiple Membership Models	1357
68.7.2	Spatial Models	1358
68.8	Conclusion	1360
	References	1361

M. Arcaya • S.V. Subramanian (✉)

Department of Society, Human Development and Health, Harvard School of Public Health,
Harvard University, Boston, MA, USA

e-mail: marcaya@hsph.harvard.edu; svsubram@hsph.harvard.edu

Abstract

Describing area-based differences in health outcomes has a long history (Kawachi and Berkman Neighborhoods and health. Oxford University Press, Oxford, 2003), and evidence of ecologic variations in health sparks interest from multiple perspectives. In particular, researchers investigate these ecologic variations for surveillance and monitoring of health disparities (Krieger et al. Am J Public Health 95:312–323, 2005) and to understand the impacts contexts have on individuals. In the latter category, causal questions motivated by evidence of area-based differences in health include the following: How much do contexts, such as neighborhoods, impact health? What is the impact of a specific contextual exposure on health? How do contexts mediate the effects of individual-level health risk factors? Ecologic factors may have tremendous importance for population health (Kawachi and Berkman Neighborhoods and health. Oxford University Press, Oxford, 2003), underscoring the value of recognizing opportunities and methodological challenges for causal inference when ecological variations in health are present. We address these issues as follows: we begin by identifying what constitutes a multilevel data analysis and present a discussion on how a range of data structures that are observed in the real world, or due to sampling design, can be accommodated within a multilevel framework. We discuss the types of research questions that typically motivate multilevel analyses and contrast the application of multilevel methods against other approaches for answering such questions with an emphasis on causal inference. After laying down the substantive motivation to utilize multilevel methods, key statistical models are specified with a description of the properties of each model. We close by presenting extensions to the basic multilevel model that allow us to incorporate realistic complexity in our analyses.

68.1 Introduction

Where you live makes a difference to your health over and above who you are (Berkman and Kawachi 2000; Jones and Moon 1993). People's lives are lived in different settings, including neighborhoods, worksites, and schools, as well as in larger scale political or economic contexts, including states, metropolitan regions, or countries. Notwithstanding individual influences on health, researchers are increasingly emphasizing the role that contexts play in shaping health and health inequities in the population. The term "multilevel" refers to the distinct levels or units of analysis, which usually, but not always, consist of individuals (at the lower level) who are nested within contextual/aggregate units (at the higher level). Viewing factors that affect health as simultaneously operating at the level of individuals and at the level of contexts is quintessentially a multilevel analytic perspective. **Multilevel methods**, vital for applying multilevel frameworks to empirical questions, consist of statistical procedures that are pertinent when:

- The observations that are being analyzed are correlated or clustered along spatial, geographic/political, and/or temporal dimensions.
- Causal processes are thought to operate simultaneously at more than one level.
- There is an intrinsic interest in describing the variability and heterogeneity in the population, over and above the focus on average relationships (Diez Roux 2002; Subramanian 2004a).

Multilevel methods are specifically geared toward the statistical analysis of data that have a *nested* structure. The nesting, typically, but not always, is hierarchical. For instance, a two-level structure would have many level 1 units nested within a smaller number of level 2 units. In educational research, the field that provided the impetus for multilevel methods, level 1 usually consists of pupils who are nested within schools at level 2. Such structures arise routinely in health and social sciences, such that level 1 and level 2 units could be workers in organizations, patients in hospitals, or individuals in neighborhoods, respectively. Third levels and higher may also be considered, for example, individuals at level 1 nested neighborhoods at level 2, which in turn are nested in cities at level 3.

The existence of **nested data structures** is neither random nor ignorable; neighborhoods and other contexts differ just as individuals do. Differences among these contexts could reflect differences among individuals nested in them, or they may arise for reasons less strongly associated with the characteristics of the individuals they encompass (e.g., children within a school may be arbitrarily assigned to classrooms). Regardless, once contexts are established, they tend to become differentiated and therefore meaningful, even if their establishment is random. This would compel analysts interested in causal questions to consider covariates at multiple levels. Researchers must *a priori* specify why they think that there will be variation in the outcome at the levels they include over and above variation at level 1. For example, do we expect variation in a given outcome at the level of small neighborhoods (e.g., census blocks) or larger neighborhoods (e.g., census tracts)? Will states additionally play a role?

Theory and subject matter knowledge must drive the decision of what levels to consider and omit, and a proper application of methods is needed to ensure valid causal inferences. This chapter is methodologically focused with an aim to help researchers identify multilevel data structures, causal questions that may be asked using a multilevel framework, substantive differences between multilevel and other approaches to handling nested data, and opportunities and challenges for causal inference at the ecologic level using multilevel approaches.

68.2 Recognizing Multilevel Data Structures

Multilevel data structures dominate observational datasets and even some experimental designs. While multilevel methods are commonly employed to make inferences about how contexts affect people, it is important to note that multilevel methods are applicable in a range of other settings as well. One well-recognized source of nested data structures is sampling design. For instance, multistage

sampling designs are often employed during large-scale survey data collection for reasons of cost and efficiency. A national population survey, for example, might involve a three-stage design, with regions sampled first, then neighborhoods, and then individuals. A design of this kind generates a three-level hierarchically clustered structure of individuals at level 1 nested within neighborhoods at level 2, which in turn are nested in regions at level 3. Observations taken on individuals living in the same neighborhood can be expected to be more alike than they would be if the sample were truly random, meaning that data points are not truly independent. Similar dependence can be expected for data collected on neighborhoods within a region.

Much documentation exists on measuring this “**design effect**” and correcting for it. Indeed, clustered designs (e.g., individuals at level 1, nested in neighborhoods at level 2, and nested in regions at level 3) are often a nuisance in traditional analysis. This “nuisance perspective,” while often appropriate to handle clustering induced by sampling design, is in contrast with the idea that individuals, neighborhoods, and regions can be seen as distinct structures that exist in the population that should be measured and modeled. Despite these differences in motivations, multilevel methods are applicable in both types of settings.

The idea of multilevel structure can also be recast, with great advantage, to address a range of other circumstances where one may anticipate clustering. Outcomes as well as their causal mechanisms are rarely stable and invariant over time, producing data structures that involve repeated measures. **Repeated measures** over time taken on a higher-level unit such as an individual or context can be considered a special case of multilevel clustered data structures. Consider the “repeated cross-sectional design” that can be structured in multilevel terms with neighborhoods at level 3, year/time at level 2, and individuals at level 1. In this example, level 2 represents repeated measurements on the neighborhoods (level 3) over time. Such a structure can be used to investigate what sorts of individuals and what sorts of neighborhoods have changed with respect to the outcome. Alternatively, there is the classic “longitudinal or panel design” in which level 1 is the measurement occasion, level 2 is the individual, and level 3 is the neighborhood. This time, the individuals are repeatedly measured at different time intervals so that it becomes possible to model changing individual behaviors within a contextual setting of, say, neighborhoods.

“Multivariate” multilevel data structures occur when different responses or outcomes reported for a higher-level unit are correlated. For example, level 1 could be a set of response variables measured on individuals at level 2 nested in neighborhoods at level 3. The “**multivariate responses**” could be, for instance, different aspects of, say, health behavior (e.g., smoking and drinking). In addition, such responses could be a mixture of “quality” (do you smoke/do you drink) and “quantity” (how many/how much), producing “mixed multivariate responses.” The substantive benefit of this approach is that it is possible to assess whether different types of behavior and whether the qualitative and quantitative aspects of each behavior are related to individual characteristics in the same or different ways. This structure also lets us study whether neighborhoods suffer or enjoy similar

levels of related characteristics, such as whether neighborhoods with a high prevalence of smoking also suffer high rates of drinking and drug use.

While the previous examples are strictly hierarchical, meaning that each level 1 observation belongs to just one level 2 unit, data structures could alternatively be nonhierarchical. For example, a model of health behavior could be formulated such that behavioral information, such as exercise frequency, is collected on individuals (level 1) who are members of both residential neighborhoods *and* workplaces at level 2. In this example, individuals are not strictly nested but rather “cross-classified” in two different higher-level groupings. Individuals are then allowed to occupy more than one set of contexts, each of which may have an important influence. For instance, individuals in a single workplace may come from many different neighborhoods, and individuals in a given neighborhood may each work in a different worksite.

A related nonhierarchical structure may occur even when there is only one type of level 2 context. For example, level 1 observations, such as students, often belong to more than one level 2 unit, such as classrooms, even if information about additional *types* of contexts is not available. In contrast to the **cross-classified structures** described above, where information is available on two types of contexts (i.e., neighborhoods and workplaces), “**multiple membership designs**” accommodate level 1 units belonging to more than one level 2 unit of the same type. A college student taking several courses can simultaneously belong to several classrooms even if we know nothing about his/her neighborhood or worksite, for example, with the contribution of each classroom weighted in relation time spent by the student in that environment. Similar arrangements can be made for individuals within neighborhoods, where weights might be determined by spatial distance or time allocation, or any other metric that is meaningful for the research question at hand. In summary, between some combination of hierarchical structures, cross-classified nesting, and multiple membership, a great complexity that is imprinted either explicitly or implicitly in data can be incorporated via multilevel models.

68.3 Causal Research Questions That Motivate Multilevel Analyses

As evidenced by the previous section, multilevel data structures are more a rule than an exception in epidemiologic research and should be considered for both technical and substantive reasons. Particularly when trying to understand observed ecologic variations in health, a wide range of causal questions are best answered using multilevel methods, and a subset of these can *only* be answered using a multilevel approach. Among the questions that may be asked and answered using a multilevel framework are the following: What is the impact of a specific contextual exposure on an outcome? How much do contexts, generally speaking, matter for a given outcome? How do contexts mediate the effects of individual-level factors?

68.3.1 Specific Ecologic Effects

Beginning with the first of these questions, assessing the effect of a contextual-level exposure on individuals often calls for a multilevel approach. To put this in perspective, we consider an influential study of progress among primary school children from the 1970s. Bennett (1976) used a single-level multiple regression analysis to claim that children exposed to “formal” style of teaching exhibited more progress than those who were not. While recognizing individual children as units of analysis, the study ignored the grouping of children into classrooms. In what was the first important example of multilevel analysis using social science data, Aitkin (Aitkin et al. 1981) reanalyzed the data and demonstrated that when the analysis accounted properly for the grouping of children (level 1) into classrooms (level 2), the progress of formally taught children could not be shown to significantly differ from the others.

Why was the initial analysis flawed? Children within any one classroom tended to be similar in their performance simply because they were taught together, thereby providing much less information than would have been the case if the study relied on a sample of children that had been taught separately. More formally, the individual samples (in this case data collected on children) were *correlated* or *clustered*. Such clustered samples do not contain as much information as simple random samples of similar size. As was shown by Aitkin (Aitkin et al. 1981), ignoring this autocorrelation, or clustering, resulted in an increased risk of finding differences and relationships where none existed.

In his investigation of student progress in 1976, Bennett was attempting to understand a *specific ecologic effect*, specifically the role of formal teaching style on children’s progress. The term *specific ecologic effect* refers to the impact a particular characteristic of a context has on lower-level units. This can be contrasted with the concept of a *common ecologic effect*, or the overall impact a higher-level environment has on lower-level units. Examples of specific ecologic effects include the impact of neighborhood poverty on individual smoking behavior, the effect of workplace breastfeeding policies on individual breastfeeding behavior, or how school-level physical education policies impact student BMI. In each of these examples, the target of inference is a quantifiable aspect of a given context, not the impact of the context generally speaking. However, researchers may wish to focus on overall contextual effects without specifying what it is about contexts that are beneficial or harmful for a given outcome.

68.3.2 Common Ecologic Effects

In contrast to the illustrations above, examples of *common ecologic effects* include the degree to which school districts influence high school seniors’ chances of going to college above and beyond individual factors, or the extent to which physician practice networks rather than patient-level factors determine if sick individuals are

prescribed antibiotics. In these examples, there are no quantifiable characteristics of the school district or physician practices that are under review; rather, the focus is on whether membership in a context matters in general and to what extent. Generic neighborhood, or contextual, effects are useful for two reasons. First, recognizing that ecologies matter is often a first step toward identifying a specific influential contextual characteristic of interest for a given outcome. Second, contextual influences on health are likely to be mediated through multiple and synergistic interactions of several neighborhoods or other contextual elements, as opposed to a single exposure. As such, common ecologic effects may be important in an intrinsic sense (Sampson et al. 2002). As opposed to a *specific ecologic effect*, which describes the impact of changing an isolated aspect of a context, a *common ecologic effect* would include the consequences of being placed in a different environment with different opportunity structures as well as different neighbors, classmates, or residents, rather than just a different value for one specific contextual characteristic.

As applied to neighborhoods and health research, for example, the basic causal question asked when studying a specific ecologic effect therefore is this: If one were to change specific neighborhood characteristic(s) of the neighborhood setting but change nothing about the individual in question, what would be the impact on the individual's health, on average? The basic causal question asked when studying common ecologic effects is to what extent does membership in a neighborhood impact an individual's health above and beyond individual factors.

It is important to emphasize that these types of causal research questions (i.e., about common ecologic effects) asked under a multilevel framework focus on variability rather than means. Consequentially, researchers are not primarily interested in effect estimates associated with specific exposures at either level 1 or level 2. It is the norm in epidemiology to investigate causation as changes in group means, although many variables of interest may cause a change in the variance of the distribution of the dependent variable and not cause a change in the mean. In fact, there has been little interest to date in variance that underlies averages (Braumoeller 2006). This is because variance is often considered a measure of uncertainty or a troublesome entity, rather than a source of substantive information even in multilevel investigations (Roux 2008). To ignore information discernible by assessing measures of variance is certainly a missed opportunity (Kawachi and Berkman 2003; Larsen and Merlo 2005; Subramanian 2004b). Researchers should always remember that the goal of social medicine is not only to increase the (mean) health of the population, but also to decrease **inequities in health** (variance).

Rather than choose, researchers can also integrate studies of specific and common ecologic effects using a multilevel framework, estimating both the effect of a specific neighborhood characteristic on an outcome and whether the neighborhood predicts the outcome in a model conditional on that neighborhood characteristic. For example, we could estimate the effect of fast-food outlet density of a neighborhood on BMI and ask whether there are additional differences in BMI across neighborhoods of similar fast-food density.

68.3.3 Composition Versus Context

Whether focused on specific or common ecologic effects, a main goal for causal inference is disentangling different sources of variation in the outcome at different levels. Evidence of variation in poor health among different neighborhoods, for instance, could be due to factors that are intrinsic to the neighborhoods themselves and are measured at the neighborhood level. In other words, the variation may be due to what are called *contextual* or *neighborhood effects*. Alternatively, variations between neighborhoods may be *compositional* or simply reflective of the characteristics of residents. For example, if people who are likely to be in poor health due to their individual characteristics are clustered in neighborhoods, we will see ecologic variations in health due to the composition of neighborhoods rather than any causal effect neighborhood exerts on residents. If we were to find that average BMI tends to be higher in neighborhoods with a higher proportion of impoverished residents, such an observation does not, by itself, provide insight into the causal question of interest, that is, does living in high-poverty neighborhoods increase individual residents' BMI compared with living in low-poverty neighborhoods? The issue, therefore, is not whether variations between different neighborhoods exist (they usually do), but what is the primary source of these variations. Put simply, are there significant contextual differences in health between neighborhoods, after taking into account the individual compositional characteristic of the neighborhood? The notions of contextual and compositional sources of variation have general relevance, and they are applicable whether the context is administrative (e.g., political boundaries), temporal (e.g., different time periods), or institutional (e.g., schools or hospitals). Multilevel studies allow us to examine such questions by considering both individual and contextual characteristics. For causal inference, it is important that variables at these levels are not completely confounded.

68.3.4 Describing Contextual Heterogeneity

Yet another motivation to employ a multilevel approach is to understand how the effect of an individual-level exposure varies according to context or how contextual effects vary according to individual factors. As an example of the former, the effect of low personal income on health may be mitigated or exacerbated by neighborhood conditions generally or by specific contextual-level factors such as presence of affordable housing, transit access, or food environment. As for the latter, individual age or sex may affect the impact vulnerability to the negative effects of neighborhood-level violence.

Describing such *contextual heterogeneity* is an important aspect of multilevel analysis and can have two interpretative dimensions. First, there may be a *different amount* of neighborhood variation for a given outcome, such that, for example, for high-income individuals, it may not matter in which neighborhoods they live (thus a lower between-neighborhood variation in health outcomes), but it matters a great

deal for the low-income residents and as such shows a large between-neighborhood variation in health outcomes. Second, there may be a *differential ordering*: neighborhoods that are high on a value for one group are low on the same value for the other and vice versa. Stated simply, the multilevel analytical question is whether the contextual neighborhood differences in poor health, after taking into account the individual composition of the neighborhood, different for different types of population groups.

In one empirical demonstration of this approach, researchers involved in the MONICA project on blood pressure investigated contextual effects on the individual-level factors of BMI and antihypertensive medication use (Merlo et al. 2004). Researchers found that contextual effects were particularly strong in overweight women on antihypertensive medication. Approximately 20 % of the individual differences in blood pressure in this sample could be attributed to a contextual effect, while only about 8 % of variability could be attributed to context in the population overall. Such a large difference may reflect disparities in the effectiveness of individual antihypertensive treatment across different national health-care systems. These findings illustrate that contextual differences may be complex such that effects may not be the same for all types of people.

68.3.5 Characterizing and Explaining Contextual Variations

Contextual differences, in addition to people's characteristics, may also be influenced by the different characteristics of neighborhoods. In other words, individual differences may interact with context, and ascertaining the relative importance of individual and neighborhood covariates is another key aspect of a multilevel analysis. For example, over and above income (individual characteristic), health may depend upon the poverty levels of neighborhoods (neighborhood characteristic), as stated previously. We might consider that the **contextual effect** of poverty is the same for both the high- and low-income individuals, suggesting that while neighborhood poverty could explain the prevalence of poor health in a poor neighborhood, it does not influence the social class inequalities in health. On the other hand, the contextual effects of poverty may be different for different groups, such that neighborhood poverty adversely affects the low-income residents, but benefits the health of high-income residents. Thus, neighborhood-level poverty not only may be related to average health achievements but also shapes social inequalities in health. The analytical question of interest is whether the effect of neighborhood-level socioeconomic characteristics on health is different for different types of people.

In sum, **casual ecologic inferences** may concern changes in means or variance and direct influences of context on health or indirect influences wherein context mediates individual exposures. Below we map a variety of **study designs** against what types of questions they enable researchers to answer. We contrast multilevel against other approaches for these various types of reviews.

68.4 Multilevel Framework: A Necessity for Understanding Ecologic Effects

Figure 68.1 identifies a typology of designs for data collection and analyses (Blakely and Woodward 2000; Subramanian et al. 2007), where the rows indicate the level or unit at which the outcome variable is being measured (i.e., at the individual level (y) or the ecological level (Y)) and the columns indicate whether the exposure is being measured at the individual level (x) or the ecological level (X). The ecological level could be any context of interest such as a school, country, or city, but we will refer to it as the neighborhood level for simplicity here. Study-type (y,x) is most commonly encountered when the researcher aims to link exposure measured at the individual level to outcomes also measured at the individual level and is common in “risk factor epidemiology.” Study-type (y,x) typically ignores ecological effects (either implicitly or explicitly), essentially remaining agnostic about context.

Conversely, study-type (Y,X) – referred to as an “ecological study” – may seem intuitively appropriate for research where higher levels (for instance, neighborhoods, regions, states, schools, and so on) are the targets of interest. However, study-type (Y,X) also misses key information in that it fails to disentangle the genuinely ecological (i.e., “**contextual**”) and the aggregate (i.e., “**compositional**”) (Moon et al. 2005) and precludes the possibility of testing heterogeneous contextual effects on different types of individuals. Ecological, or contextual, effects reflect predictors and associated mechanisms that operate primarily at the contextual level. That is, context exerts a causal influence on individuals. The search for such measures and their scientific validation and assessment is an area of active research (Kawachi and Berkman 2003). Aggregate effects, or compositional effects, in contrast, equate the effect of a neighborhood with the sum of the individual effects associated with the people living within the neighborhood. In this situation, the interpretative question becomes particularly relevant. If common membership of a neighborhood by a set of individuals brings about an effect that is over and above those resulting from individual characteristics, then there may indeed be an ecological effect. Under a purely ecological study, it is impossible to differentiate between ecological, or contextual, and aggregate, or compositional effects.

Study-type (y,X) provides a multilevel approach under which an ecological exposure is linked to an individual outcome. A more complete representation would be (y,x,X) such that we have an individual outcome, individual confounders (x), and neighborhood exposure. Such data would reflect a multilevel structure of individuals nested within neighborhoods. A fundamental motivation for study-type (y,x,X) is to distinguish “neighborhood differences” from “the difference a neighborhood makes” (Moon et al. 2005). Stated differently, ecological effects on the individual outcome should be ascertained *after* individual factors that reflect the composition of the places (and may be potential confounders) have been controlled. Indeed, compositional explanations for ecological variations in health are common. It nonetheless makes intuitive sense to test for the possibility of ecological effects. Besides anticipating their impact on individual outcomes,

		Exposure	
		Individual (x) (measured at individual level)	Ecologic (X) (measured at ecological level)
Outcome	Individual (y)	(y,x) Traditional risk factor study	(y,X) Multilevel study
	Ecologic (Y)	$(Y,x)^{(A)}$	(Y,X) Ecological study

Fig. 68.1 Typology of studies (Subramanian et al. 2007). Note: ^(A) This type of study is impossible to specify as it stands. Practically speaking, it will either take the form of (Y,X) , that is, ecological study, where X will now simply be central tendency of x . Or, if dis-aggregation of Y is possible, so that we can observe y , then it will be equivalent to (y,x) (Source: Subramanian et al. 2009)

compositional factors may vary by context. Moreover, composition itself has an intrinsic ecologic dimension; the very fact that individual (compositional) factors may drive ecologic variations serves as a reminder that the real understanding of **ecologic effects** is likely to be complex.

The multilevel framework with its simultaneous examination of the characteristics of the individuals at one level and the context or ecologies in which they are located at another level accordingly offers a comprehensive framework for understanding the ways in which places can affect people (contextual) and/or people can affect places (composition). It likewise allows for a more precise distinction between **aggregative fallacy** versus ecologic effects (Subramanian et al. 2009).

68.5 Multilevel Statistical Analysis

In the presence of a multilevel data, as described in Sect. 68.2, and having motivations as discussed in Sect. 68.3, there are substantive as well as technical reasons to use **multilevel statistical models** to analyze such data (Goldstein 2003; Raudenbush and Bryk 2002). We shall not review the basic principles of multilevel modeling here as they have been described elsewhere in the context of health research (Blakely and Subramanian 2006; Kawachi and Berkman 2003; Moon et al. 2005), but rather provide a brief overview of the type of models invoked for identifying ecologic effects discussed in Sect. 68.3.

Multilevel models can broadly be expressed as follows: response = fixed/average parameters + (random/variance parameters). A multilevel model is distinguished from a conventional regression model by the random part of the model. Conventional models usually restrict the random portion of a model to a single term (called as “error terms” or “residuals”); in the multilevel regression model, the random part of the statistical model is expanded.

Suppose we are interested in studying the variation in health, represented by a “health score,” as a function of certain individual and **neighborhood predictors**. Let us assume that a researcher collected data on a sample of 50 neighborhoods and, within each of these neighborhoods, a random sample of individuals. This design produces a two-level structure where the outcome is a health score, y , for individual i in neighborhood j . Individual level, poverty, x_{1ij} , is collected and coded as 0 if not poor and 1 poor, for every individual i in neighborhood j . The researcher also collects data on one neighborhood predictor, w_{1j} , a socioeconomic deprivation index in neighborhood j .

68.5.1 Variance Component or Random Intercepts Model

We will now develop regression equations at the individual and neighborhood levels of analysis in order to understand how a multilevel model would accommodate the data structure above. In the illustration considered here, models would have to be specified at two levels, level 1 and level 2, though additional levels are possible if the research question calls for this. For example, a model at level 1 can be formally expressed as

$$y_{ij} = \beta_{0j} + \beta_1 x_{1ij} + e_{0ij} \quad (68.1)$$

In this level 1 model, β_{0j} (associated with a constant, x_{0ij} , a constant vector of 1s) is the mean health score for the j^{th} neighborhood for the nonpoor group; β_1 is the average differential in health score associated with individual poverty status (x_{1ij}) across all neighborhoods. Meanwhile, e_{0ij} is the individual, or the level 1, residual term. We create a two-level model by allowing β_{0j} to become a **random variable** as

$$\beta_{0j} = \beta_0 + u_{0j} \quad (68.2)$$

where u_{0j} is an error term that describes the random neighborhood-specific displacement associated with the overall mean health score (β_0) for the nonpoor group. Since we do not allow, at this stage, the average differential for the poor and nonpoor group (β_1) to vary across neighborhoods, u_{0j} is assumed to be same for both groups. We refer to Eq. (68.2) as the level 2, between-neighborhood model.

We can now integrate these two parts to form a multilevel model. We first substitute the level 2 model [Eq. (68.2)] into level 1 model [Eq. (68.1)] and then group the fixed and random part components of the model (bracketed). The result is the following combined, also referred to as **random intercepts** or **variance components**, model:

$$y_{ij} = \beta_0 + \beta_1 x_{1ij} + (u_{0j} + e_{0ij}) \quad (68.3)$$

Our response variable y_{ij} is the sum of fixed and random model components. Assuming a normal distribution with a 0 mean, we can estimate a variance at level 1

(σ_{e0}^2 : the between-individual within-neighborhood variation) and level 2 (σ_{u0}^2 : the between-neighborhood variation), both conditional on fixed poverty differences in health score. The structure of the random component of this model, specifically that it contains more than one residual term, distinguishes the multilevel model from the standard linear regression models or analysis of variance-type analysis. The underlying random structure (variance-covariance) of the model specified in Eq. (68.3) is

$$\text{Var}[u_{0j}] \sim N(0, \sigma_{u0}^2); \text{Var}[e_{0ij}] \sim N(0, \sigma_{e0}^2); \text{and Cov}[u_{0j}, e_{0ij}] = 0$$

This underlying **variance-covariance structure** requires special estimation procedures in order to obtain satisfactory parameter estimates (Goldstein 2003).

The model specified in Eq. (68.3) with the above random structure is typically used to partition variation according to the different levels, with the variance in y_{ij} being the sum of σ_{u0}^2 and σ_{e0}^2 . This partitioning allows us to calculate a statistic known as the **intraclass correlation**, or *intra-unit correlation*, or more generally *variance partitioning coefficient* (Goldstein et al. 2002), which represents the degree of similarity between two randomly chosen individuals within a neighborhood. When individuals within a neighborhood are nearly identical, this statistic approaches 1. When individuals are truly independent, meaning that the contextual level explains practically none of the variance among observations, the statistic approaches zero. The variance partitioning coefficient can be expressed as

$$\rho = \frac{\sigma_{u0}^2}{\sigma_{u0}^2 + \sigma_{e0}^2} \quad (68.4)$$

Note that Eq. (68.3) estimates a variance based on the observed sample of neighborhoods. While this is important to establish the overall importance of neighborhoods as a unit or level (i.e., the **common ecologic effect**), another quantity of interest may pertain to estimating whether living in neighborhood j_1 , as compared to neighborhood j_3 , for example, predicts a different health score conditional on compositional influences of covariates. Given Eq. (68.3), we can estimate for each level 2 unit:

$$\hat{u}_{0j} = E(u_{0j}|Y, \hat{\beta}, \hat{\Omega}) \quad (68.5)$$

The quantities \hat{u}_{0j} are referred to as “estimated” or “predicted” residuals, or using Bayesian terminology, as “posterior” residual estimates. To calculate these posterior residuals, we first calculate the raw residual for level 1 units from the random intercepts model:

$$r_{ij} = \hat{y}_{ij} - y_{ij} \quad (68.6)$$

The r_{ij} contains both level 1 and level 2 residuals, not separated out. We then calculate the mean of the raw residuals for each level 2 unit, giving us r_j .

Essentially, these are the estimates that we would get if we had specified neighborhoods as dummy variables in the fixed part of the model (with no intercept). These are also referred to as the unshrunken estimates of u_{0j} . We use the mean of the raw residuals for each level 2 unit, r_j , to calculate the **posterior residuals** as

$$\hat{u}_{0j} = r_j \times \frac{\sigma_{u0}^2}{\sigma_{u0}^2 + \sigma_{e0}^2/n_j} \quad (68.7)$$

where σ_{u0}^2 , σ_{e0}^2 , and r_j are defined above and n_j is the number of people within each neighborhood. This formula for \hat{u}_{0j} uses the level 1 and level 2 variances and the number of people observed in neighborhood j to scale the observed level 2 residual (r_j). As the level 1 variance declines or the sample size increases, the scale factor approaches 1, and thus, estimated \hat{u}_{0j} approaches r_j .

Neighborhood-level residuals are random variables assumed to originate from a distribution and whose parameter values quantify the variation among the higher-level or neighborhood units (Goldstein 2003). That is, observed neighborhoods are treated as a sample from a larger distribution of neighborhoods, of which the sample is representative. Another interpretation is that each \hat{u}_{0j} estimates neighborhood j 's departure from expected mean outcome. This interpretation is premised on the assumption that each neighborhood belongs to a population of neighborhoods, and the distribution of the population provides information about plausible values for neighborhood j (Goldstein 2003). For a neighborhood with only a few individuals, we can obtain more precise estimates by combining the population and neighborhood-specific observations than if we were to ignore the population membership assumption and use only the information from that neighborhood. When the estimated residuals at higher-level units are of interest in their own right, we need to provide standard errors, interval estimates, and significance tests as well as point estimates for them (Goldstein 2003).

68.5.2 Contrasting Multilevel and Other Approaches: Fixed Versus Random Effects

While this chapter has argued for the utility of multilevel methods for ecologic inference, we note that there are alternative approaches that can address nested data structures and may be appropriate under some circumstances, especially when measuring and modeling variance are not prime targets of analysis. For example, multilevel methods may not be *required* when clustering of data is viewed as a nuisance or when mean associations between a contextual variable and individual outcome are of prime interest, though multilevel models may still provide richer analytic options under these conditions. We return to examples addressed earlier to explore alternate specifications.

Multilevel methods help us understand the effects of contexts generally and/or the effects of specific contextual variables on a given outcome. However, this is not the only approach to understanding differences among contexts. Taking the

example of health variations across neighborhoods in Sect. 68.5.1, it is worth emphasizing that the “neighborhood effect,” u_{0j} , described in Eq. (68.2) can be treated in one of the two ways. Rather than treat the neighborhoods as a sample from a larger distribution of contexts, one can estimate each neighborhood separately as a fixed effect (i.e., treat them as a variable; with 50 neighborhoods, there will be 49 additional parameters to be estimated).

It is worth drawing this parallel between multilevel or a random-effects model (Eq. 68.3) and the conventional OLS or **fixed-effects regression model** because each has different strengths. Consider the fixed-effects model, whereby the neighborhood effect is estimated by including a dummy for each neighborhood, as shown below:

$$y_{ij} = \beta_0 + \beta_1 x_{1ij} + \beta N_j + (e_{0ij}) \quad (68.8)$$

where N_j is a vector of dummy variables for $N - 1$ neighborhoods. As stated above, the key conceptual difference between the fixed- and the random-effects approach to modeling neighborhoods is that while the fixed coefficients are estimated separately in multilevel modeling, the random differentials (u_{0j}) are conceptualized as coming from a distribution (Goldstein 2003). This multilevel conceptualization results in three practical benefits (Jones and Bullen 1994):

1. *Pooling information* between neighborhoods, with all the data contributing to the combined estimation of the fixed and random parts; in particular, the overall regression terms are based on the information for all neighborhoods.
2. *Borrowing strength*, whereby neighborhood-specific relations that are imprecisely estimated benefit from the information for other neighborhoods.
3. **Precision-weighted estimation**, whereby unreliable neighborhood-specific fixed estimates are differentially down weighted or shrunk toward the overall city-wide estimate. A reliably estimated within-neighborhood relation will be largely immune to this shrinkage.

The random-effects and the fixed-effects estimates for each neighborhood, meanwhile, are related (Jones and Bullen 1994). The neighborhood-specific random intercept (β_{0j}) in a multilevel model is a weighted combination of the specific neighborhood coefficient in a fixed-effects model (β_{0j}^*) and the overall multilevel intercept (β_0), in the following way:

$$\beta_{0j} = w_j \beta_{0j}^* + (1 - w_j) \beta_0 \quad (68.9)$$

with the overall multilevel intercept being a weighted average of all the fixed intercepts:

$$\beta_0 = \left(\sum w_j \beta_{0j}^* \right) / \sum w_j \quad (68.10)$$

Each neighborhood weight is the ratio of the true between-neighborhood parameter variance to the total variance, which additionally includes sampling variance

resulting from observing a sample from the neighborhood. Consequently, the weights represent the reliability or precision of the fixed terms:

$$w_j = \frac{\sigma_{uo}^2}{v_j^2 + \sigma_{uo}^2} \quad (68.11)$$

where the random sampling variance of the fixed parameter is

$$v_j^2 = \frac{\sigma_{e0}^2}{n_j} \quad (68.12)$$

with n_j being the number of observations within each neighborhood. This weighting structure means that when there are genuine differences among neighborhoods and sample sizes within neighborhoods are large, the sampling variance will be small in comparison to the total variance, and the random neighborhood effect estimate will be very close to the fixed neighborhood effect. As the sampling variance increases, however, the weight will be less than 1, and the multilevel estimate will increasingly be influenced by the overall intercept based on pooling across neighborhoods. “**Shrinkage estimates**” allow the data to determine an appropriate compromise between specific estimates for different neighborhoods and the overall fixed estimate that pools information across places over the entire sample (Jones and Bullen 1994).

Importantly, the fixed-effects approach to modeling neighborhood differences using cross-sectional data is *not* a choice for a typical multilevel research question, where there is an intrinsic interest in an exposure measured at the level of neighborhood such as the one specified in Eq. (68.3); in such instances, a multilevel modeling approach is a necessity. This is because the dummy variables associated with the neighborhoods (measuring the fixed effects of each neighborhood) and the neighborhood exposure of interest are perfectly confounded and, as such, the effect of the neighborhood-level exposure cannot be estimated. Thus, the fixed-effects specification to understand neighborhood differences is unsuitable for the sort of complex questions which multilevel modeling can address.

Assuming that multilevel, or random, rather than fixed-effects approach is chosen, we return to our original example of health scores across neighborhoods. Building on our basic multilevel model, we can also expand the *random structure* in Eq. (68.3) by allowing the fixed effect of individual poverty (β_1) to randomly vary across neighborhoods in the following manner:

$$y_{ij} = \beta_{0j} + \beta_{1j}x_{1ij} + e_{0ij} \quad (68.13)$$

At level 2, there will now be two models:

$$\beta_{0j} = \beta_0 + u_{0j} \quad (68.14)$$

$$\beta_{1j} = \beta_1 + u_{1j} \quad (68.15)$$

Substituting the level 2 models in Eqs. (68.14) and (68.15) into the level 1 model in Eq. (68.13) gives

$$y_{ij} = \beta_0 + \beta_1 x_{1ij} + (u_{0j} + u_{1j}x_{1ij} + e_{0ij}) \quad (68.16)$$

Equation (68.13) allows us to ask whether individual poverty has different consequences for health in different neighborhoods. This is one example of a research question focused on the indirect effect contexts can have on health, specifically as mediators of individual-level risk factors.

Across neighborhoods, the mean health score for nonpoor is β_0 , and $\beta_0 + \beta_1$ is the mean health score for the poor, and the mean “poverty differential” is β_1 . The poverty differential is no longer constant across neighborhoods, but varies by the amount u_{1j} around the mean, β_1 . Such models are also referred to as *random-slopes* or *random-coefficient models*. These models have a more complex variance-covariance structure than before:

$$\text{Var} \begin{bmatrix} u_{0j} \\ u_{1j} \end{bmatrix} \sim N \left(0, \begin{bmatrix} \sigma_{u0}^2 & \sigma_{u0u1}\sigma_{u1}^2 \\ \sigma_{u0u1}\sigma_{u1}^2 & \sigma_{u1}^2 \end{bmatrix} \right) \quad (68.17)$$

and

$$\text{Var}[e_{0ij}] \sim N(0, \sigma_{e0}^2) \quad (68.18)$$

With this formulation, it is no longer straightforward to think in terms of a summary **intraclass correlation** statistic ρ as the level 2 variation is now a function of a individual predictor variable, x_{1ij} . In our exemplification when x_{1ij} is a dummy variable, we will have two variances estimated at level 2: one for nonpoor which is

$$\sigma_{u0}^2 \quad (68.19)$$

and one for poor which is

$$\sigma_{u0}^2 + 2\sigma_{u0u1}x_{1ij} + \sigma_{u1}^2x_{1ij}^2 \quad (68.20)$$

That is, level 2 variation will be a “quadratic” function of the individual predictor variable when x_{1ij} is a continuous predictor. Thus, the notion of “random intercepts and slopes,” while intuitive, is not entirely appropriate. Rather, what these models are really doing is modeling variance as some function (constant, quadratic, or linear) of a predictor variable (Kawachi and Berkman 2003).

Building on the above perspective of modeling the variance-covariance function (as opposed to “random intercepts and slopes”), we can extend the concept to modeling variance function at level 1. It is extremely common to assume that the

variance is “homoskedastic” in the random part at level 1 (σ_{e0}^2); Eq. (68.16), and indeed researchers seldom report whether this assumption was tested or not. One strategy would be to model the different variances for poor and nonpoor of the following form:

$$y_{ij} = \beta_0 + \beta_1 x_{1ij} + (u_{0j} + u_{1j}x_{1ij} + e_{1ij}x_{1ij} + e_{2ij}x_{2ij}) \quad (68.21)$$

where $x_{1ij} = 0$ for nonpoor, 1 for poor; the new variable $x_{2ij} = 1$ for nonpoor, 0 for poor, with $Var[e_{1ij}] = \sigma_{e1}^2$ giving the variance for poor and $Var[e_{2ij}] = \sigma_{e2}^2$ giving the variance for nonpoor; and $Cov[e_{1ij}, e_{2ij}] = 0$. There are other parsimonious ways to model level 1 variation in the presence of a number of predictor variables (Kawachi and Berkman 2003). With this specification, we do not have an interpretation of the random level 1 coefficients as “**random slopes**” as we did at level 2. The level 1 parameters σ_{e1}^2 and σ_{e2}^2 describe the complexity of level 1 variation, which is no longer homoskedastic (Goldstein 2003). Anticipating and modeling heteroskedasticity or heterogeneity at the individual level may be important in multilevel analysis as there may be **cross-level confounding** – what may appear to be neighborhood heterogeneity (level 2) to be explained by some ecological variable could be due to a failure to take account of the between-individual (within neighborhood) heterogeneity (level 1).

Such exploration of variance is not possible using a fixed-effects approach, though fixed-effects models have their own merits. A fixed-effects approach may be appropriate if researchers are interested in making inferences about specific neighborhoods in a dataset, though much is lost in terms of flexibility and exploiting the richness of heterogeneity that underlies a dataset. A single-level model is used to control stringently for confounding by shared neighborhood environment such that main effects apart from neighborhood environment can be estimated. A fixed-effects approach also allows for direct comparisons among places, as an effect on the outcome of interest will be estimated for each neighborhood. On the other hand, if neighborhoods are treated as a (random) sample from a population of neighborhoods (which might include neighborhoods in future studies if one has complete population data), the target of inference is the variation between neighborhoods in general. Adopting this multilevel statistical approach makes u_{0j} a random variable at level 2 in a two-level statistical model.

The degree to which stringent control for potential confounding by neighborhood is needed and the value of directly comparing specific neighborhoods to answer the research question at hand should drive the choice of a fixed-effects versus random-effects approach. When **specific ecologic effects** are of interest, an attractive feature of multilevel models – one that is perhaps most commonly used in social science research – is their utility in modeling neighborhood *and* individual characteristics, and any interaction between them, simultaneously. We will consider the underlying level 2 model related to Eq. (68.20), which is exactly the same

as specified in Eqs. (68.14)–(68.15), but now including a level 2 predictor, w_{1j} , the deprivation index for neighborhood j :

$$\beta_{0j} = \beta_0 + \alpha_1 w_{1j} + u_{0j} \quad (68.22)$$

$$\beta_{1j} = \beta_1 + \alpha_2 w_{1j} + u_{1j} \quad (68.23)$$

Note that the separate specification of micro- and macro-models correctly recognizes that the contextual variables (w_{1j}) are predictors of between-neighborhood differences. The extension of Eq. (68.21) will now be

$$y_{ij} = \beta_0 + \beta_1 x_{1ij} + \alpha_1 w_{1j} + \alpha_2 w_{1j} x_{1ij} + (u_{0j} + u_{1j} x_{1ij} + e_{1ij} x_{1ij} + e_{2ij} x_{2ij}) \quad (68.24)$$

The combined formulation in Eq. (68.24) highlights an important feature, the presence of an interaction between a level 2 and level 1 predictor ($w_{1j} x_{1ij}$), represented by the fixed parameter, α_2 . Now, α_1 estimates the marginal change in health score for a unit change in the neighborhood deprivation index for the nonpoor, and α_2 estimates the extent to which the marginal change in health score for unit change in the neighborhood deprivation index is *different* for the poor. This multilevel statistical formulation allows ***cross-level effect modification or interaction*** between individual and neighborhood characteristics to be robustly specified and estimated.

Multilevel models are concerned with modeling both the average and the variation around the average, at different levels. To accomplish this, they consist of two sets of parameters: those summarizing the average relationships(s) and those summarizing the variation around the average at both the level of individuals and neighborhoods. Models presented in preceding section can be easily adapted to other structures with nesting of level 1 units within level 2 units. Additionally, these models can be extended to three or more levels. While the preceding discussion considered a single normally distributed response variable for illustration, multilevel models are capable of handling a wide range of responses. These include binary outcomes, proportions (as logit, log-log, and probit models), multiple categories (as ordered and unordered multinomial models), and counts (as Poisson and negative binomial distribution models). In essence, these models work by assuming a specific, “non-Gaussian” distribution for the random part at level 1 while maintaining the normality assumptions for random parts at higher levels. Consequently, the discussion presented in this entry focusing at the neighborhood level would continue to hold regardless of the nature of the response variable, with some exceptions. For instance, determining intraclass correlation or partitioning variances across individual and neighborhood levels in complex nonlinear multilevel logistic models is not straightforward (see elsewhere for details (Goldstein et al. 2002; Browne et al. 2005)). **Fixed-effects models**, by contrast, are not suited for such analyses, though they are appropriate for some research questions.

68.5.3 Contrasting Multilevel and Other Approaches: Marginal Versus Mixed Approaches

Aside from a purposeful focus on variability within a dataset, a primary motivation for applying multilevel methods to nested data is to correct for the nonindependence of level 1 observations within level 2 units. To explore this motivation, we return to the example of Aitkin's work in assessing the effect of teaching style on student progress given in Sect. 68.3.1. To answer the question of whether formal teaching style was associated with student progress on average across the population, a marginal (also known as a "population average" or **generalized estimation equation** [GEE]) model could account for clustering of students within classes and provide a "correct" answer. In fact, a marginal model specification would be less sensitive to the assumptions of random effect as compared to mixed or multilevel models (Raudenbush and Bryk 2002). However, under a marginal approach, valuable information would be lost. **Marginal models**, which do not attempt to model data structure in terms of correlations and variances, are only able to provide information about whether, on average, there is an association between two variables. Marginal models are not designed to answer other questions, such as whether the effect of an individual factor varies across contexts, or the extent to which contexts matter for an outcome. Yet, showing the average association between a contextual attribute and individual outcome is only one of several ways to consider ecologic effects. Complex heterogeneity can underlie marginal associations and may also be of interest to researchers. Multilevel approaches both account for the effect of clustering on population average estimates and can capture additional and valuable information on variance. Semiparametric specifications of random-effects distributions let the data determine the distribution for random effects, thereby reducing the reliance of the model on parametric assumptions and making inferences based on the **mixed model** more robust (Subramanian and O'Malley 2010).

68.6 Sources of Bias in Ecologic Inference

Although multilevel studies have much to offer in terms of understanding ecologic effects and heterogeneity, it is important to keep in mind several sources of **bias** that may arise in studies of contextual effects. The problem of bias is not unique to **multilevel approaches**, but rather to observational data in general, which are most often the subjects of analysis by multilevel methods. It is therefore worth noting common pitfalls of trying to identify average treatment effects using multilevel models to analyze observational data. To ground this discussion, we consider various ways in which health might vary across neighborhoods, though these concerns are applicable to other contexts and units of observation as well.

68.6.1 Reverse Causation

“**Reverse causation**” is at play if our outcome of interest has actually preceded and caused our exposure of interest rather than the other way around. Imagine a research study aimed at uncovering whether the presence of parks in a neighborhood makes residents healthier finds that neighborhood park density and fitness are indeed related. If this relationship were the result of healthy people moving to park-filled neighborhoods, perhaps because they are fit enough to take advantage of jogging trails, we would not want to attribute resident health to a contextual effect of park exposure. Rather, the relationship would imply reverse causation, whereby the outcome status drives the extent to which an exposure is experienced. In such a situation, an effect estimate describing how neighborhood-level park availability impacts fitness would be biased.

68.6.2 Uncontrolled Confounding

When neighborhood-level exposures and individual outcomes are driven by the same factors and these factors cannot be controlled, confounding will bias average treatment effect estimates. For example, if resident preferences in food drive both individual food choices, and therefore BMI, and neighborhood-level fast-food density measures, confounding limits our ability to understand the causal effect of fast-food density in a neighborhood on BMI.

68.6.3 Covariate Balance

A lesser known source of bias in contextual studies is the misspecification of covariates. This can occur when distributions of covariates in different neighborhoods do not overlap or do overlap very little. For example, consider a study of individual smoking predicted by neighborhood violence. Imagine that individual income is associated with living in a neighborhood that experiences a lot of violence and income also affects the risk of smoking. If the distribution of income in violent versus safer neighborhoods is such that no rich people lived in high-violence communities and no poor people lived in safe neighborhoods, then it is impossible to control for income when estimating the effect of neighborhood violence on smoking. This is because only extrapolated data exist that would allow us to estimate the effect of violence conditional on income.

68.6.4 Treatment Heterogeneity

Finally, average treatment effect estimates may be biased when treatment effects are heterogeneous. That is, an exposure may have one effect in a certain population

and an opposite, or at least meaningfully different, effect in another. Imagine trying to estimate the effect of a sun exposure on skin cancer if the only people willing to live in sunny environments were those who were minimally susceptible sun burn. Anyone who felt they were susceptible to sunburn would choose to live in areas without much sun exposure under this scenario. Were residents organized in this way, a study of sun exposure and skin cancer would conclude that the two variables were not related. In this scenario, while the average treatment effect would be biased, the effect of “treatment” (i.e., sun exposure) on the “treated” (i.e., those living in sunny areas) would be a valid estimate. The plausibility of heterogeneous treatment effects is specific to the social and biological background for each research question. Furthermore, such “**treatment heterogeneity**” can be of substantive importance in its own right.

68.6.5 Alternatives to Observational Data

While these sources of bias are not unique to multilevel data structures, they are common pitfalls to be aware of when making ecologic inference via multilevel methods. Concerns that observational data collected on people in contexts can never be completely free of endogeneity, or confounding by preference, have led some researchers to call for randomized community trials as the only valid way of estimating neighborhood effects (Oakes 2004).

While there are multiple flaws in this thinking (Subramanian 2004b), two types of experiments can be conceptualized for identifying ecologic effects. One approach randomly assigns individuals to neighborhoods under experiments called “mobility programs.” The Gautreaux (Rosenbaum 1995) and Moving to Opportunity (MTO) (Katz et al. 2000) studies provide the only experimental investigation of **neighborhood effects** of socioeconomic and health outcomes (although Gautreaux was technically quasi-experimental) under this type of scheme. As background, MTO randomized families in public housing to receive housing vouchers for use on the private market with different stipulations attached or to receive no new assistance. Some families received vouchers that had to be used in low-poverty neighborhoods, while others were able to use their vouchers without geographic restrictions. Follow-up on the families, who were moved between 1994 and 1997, shows that randomization to low-poverty neighborhoods decreased rates of obesity and diabetes and mental health problems in adults (Katz et al. 2000; Ludwig et al. 2011).

This study provides strong evidence that neighborhoods matter for health, though it is unclear whether MTO results can be interpreted as specific or **common ecologic effects**. While the study was designed to review the effects of neighborhood poverty on various aspects of well-being (a specific ecologic effect), the impact of poverty cannot be isolated if individuals have ties to their neighborhoods that influence well-being through means other than contextual percent poverty. Such interventions are sometimes called “fat hand” because the clumsy intervening hand has altered more than the specific causal agent of interest. Similarly, it is

difficult to impossible to intervene on one neighborhood characteristic without affecting others. For example, interventions that seek to lower levels of community violence would almost always also impact levels of community trust, aspects of the built environment, and other factors. However, if it is impossible to change one neighborhood characteristic without changing others, the causal effect of isolated interventions may not be of much interest.

A second type of experiment for identifying ecologic effects involves randomly assigning neighborhoods various characteristics, by, for example, randomly choosing areas in which bike lanes will be striped or parks enhanced. There are two challenges with these types of interventions, which rely on pre- and post-intervention data points to assess impacts. First, secular trends must be accounted, meaning “control” neighborhoods that do not receive an intervention are also followed. Secondly, required sample sizes are quite high. This is because the number of level 2 units, in this case neighborhood, rather than level 1 units, in this case residents, drives the **effective sample size**. Random assignment of potentially meaningful treatments to hundreds, if not thousands of neighborhoods, depending on the anticipated magnitude of the effect, would be a political and fiscal challenge, to say the least. To this end, those interested in estimating ecologic effects should be trained to analyze experimental, quasi-experimental, and observation data collected on multiple levels of analysis.

68.7 Extensions to the Basic Multilevel Models

68.7.1 Multiple Membership Models

While multilevel methods are powerful and flexible tools for studying ecologic effects, current implementations of multilevel models have generally failed to exploit the full capabilities of the analytical framework (Moon et al. 2005; Subramanian 2004a). Much, if not all, of the current research linking neighborhoods and health is cross-sectional and assumes a hierarchical structure of individuals nested within neighborhoods. This simplistic scenario ignores the possible data structures discussed in this chapter, including that an individual might move several times and as such reflect neighborhood effects drawn from several contexts or that other competing contexts (e.g., schools, workplaces, hospital settings) may simultaneously contribute to contextual effects. Figure 68.2 provides a visual illustration of one complex, but realistic multilevel structure for neighborhoods and health research, where time measurements (level 1) are nested within individuals (level 2) who are in turn nested within neighborhoods (level 3). Importantly, individuals are assigned different weights for the time spent in each neighborhood. For example, individual 25 moved from neighborhood 1 to neighborhood 25 during the time period t1–t2, spending 20 % of his/her time in neighborhood 1 and 80 % in his/her new neighborhood. This **multiple membership design** would allow control of changing context as well as changing composition. Such designs could be extended to incorporate memberships to additional contexts, such as workplaces or schools.

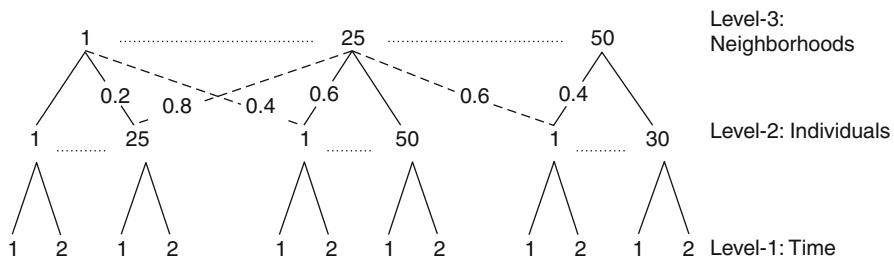


Fig. 68.2 Multilevel structure of **repeated measurements** of individuals over time across neighborhoods with individuals having multiple membership to different neighborhoods across the time span (Source: Subramanian 2004a)

It can also be extended to consider the effects of proximate contexts, weighted according to distance or other spatial measures. So, for example, the geographic distribution of disease can be seen not only as a matter of composition and the immediate context in which an outcome occurs but also a consequence of the impact of nearby contexts with nearer areas being more influential than more distant ones. This is also called **spatial autocorrelation** and forms an important area of spatial statistical research. While such analyses require high-quality longitudinal and context-referenced data, models that incorporate such “realistic complexity” (Best et al. 1996) are likely to improve our understanding of true neighborhood effects. While the foregoing discussion provides a sound rationale to adopt a multilevel analytic approach for modeling ecologic effects, they obviously do not overcome the limitations intrinsic to any observational study design, single level or multilevel.

68.7.2 Spatial Models

As noted above, the geographic distribution of disease can be seen as a consequence of the impact of nearby contexts with nearer areas being more influential than more distant ones. This idea that proximate observations and contexts are correlated by virtue of their locations in space drives the application of multilevel methods for the analysis of spatial data. Just as we can nest observations in more than one context, such as people in multiple workplaces or neighborhoods through **cross-classified** and multiple membership models, we can also nest observations within geographically defined “spatial patches” rather than within administrative boundaries. That is, rather than assign some observation higher-level groupings based on membership, we can base assignment in higher levels on space. To do this, a list of neighbors must be constructed for each observation. To know who is a neighbor, information about space is needed.

We can construct a **weights matrix** that defines the spatial relationships among all observations in a dataset to identify their spatial relationship. For observations considered “neighbors” estimates are allowed to be correlated, just as observations

Fig. 68.3 Conceptual map of nine areas a, \dots, I

a	b	c
d	e	f
g	h	i

	a	b	c	d	e	f	g	h	i
A	0	1	0	1	1	0	0	0	0
B	1	0	1	1	1	0	0	0	0
C	0	1	0	0	1	1	0	0	0
D	1	1	0	0	1	0	1	1	0
E	1	1	1	1	0	1	1	1	1
F	0	0	1	0	1	0	0	1	1
G	0	0	0	1	1	0	0	1	0
H	0	0	0	1	1	1	1	0	1
I	0	0	0	0	1	1	0	1	0

Fig. 68.4 First order queen-based contiguity matrix for area a, \dots, I

that share a context are allowed to be correlated in traditional hierarchical models. That is, the assumption that level 2 residuals, u_{0j} , are spatially unstructured can be varied, for example, such that u_{0j} is allowed to follow a conditional autoregressive prior based on neighborhood contiguity. Non-neighbors are assumed to be independent, unless otherwise specified at other levels. A common neighborhood structure is “queen-based contiguity.” The term “queen-based contiguity” derives from chess where the queen can move in any direction from her starting location and implies that all shared borders constitute first-order neighbors (Fischer and Getis 2010). To illustrate this scheme, imagine a given map of nine areas a, \dots, i shown in Fig. 68.3.

Using a first-order regional structure using queen-based contiguity, this map translates into a 9×9 adjacency matrix, as shown in Fig. 68.4.

Pairs of adjacent observations receive a value of 1, and all other pairs are assigned 0. To look up the relationship between subjects a and b , we find observation a in the first row and observation b in the second column. Reading across the matrix’s first row, the second cell contains 1 because a is adjacent to b on the map. By contrast, the cell indicating the relationship between a and c is a 0 (first row, third column) because the two areas are not adjacent on the map.

Of course, first-order adjacency matters for only some research questions, while different conceptualizations are more appropriate in other settings. Testing for and understanding spatial clustering require hypotheses of how relevant processes function over space. Contagion and/or influence by spatially patterned variables may be structured by distance, adjacency, shared access to transportation networks,

or other factors, and each of these constructs can be translated into a **spatial weights matrix**. Researchers should choose a spatial weights matrix based on expert knowledge of the topic under study. Below are other commonly employed specifications.

Distance. Distance should be considered for processes that literally operate through space. For example, studying a disease spread by insect vectors that can only travel up to 800 m from their nests, researchers might create a spatial weights matrix where all pairs of points within 800 m of each other are neighbors and all others are not.

Spatial relationships could alternatively be described by the distance between the two points where cell values would simply be the number of meters between point pairs. To provide further flexibility, these distances could be calculated either in Euclidian terms, along a road network, or by another scheme (Apparicio et al. 2008). **Geographic information systems** will also allow users to calculate “cost distances” between points that can account for the slope of a walking route, cost of tolls on a road, or even total trip time so that distances could be conceptualized in terms of minutes commuting or gasoline costs.

Adjacency. Where shared borders or regionalism matters (e.g., immigration or trade), adjacency may be a more useful concept. Adjacency can be conceptualized flexibly (Fischer and Getis 2010), where the presence, amount, or even spatial orientation of a shared border is used to construct a spatial weights matrix. Further, second-order and higher neighbors (i.e., a neighbor’s neighbor) can be ignored or included.

Nearest Neighbors. A hybrid approach between the distance and adjacency approaches is to select k nearest neighbors, with proximity based on centroid distances, from a pool of adjacent areas for each subject.

Regardless of neighborhood specification, **conditional autoregressive (CAR) models** can be fit to spatial data using multilevel methods. This is because CAR models are simply extensions of traditional original random-effects models. Instead of nesting a given observation completely within one higher-level unit, CAR models allow observations to be “cross-classified,” or influenced by a *group* of neighbors, according to any spatial weights matrix such as those described above.

Bayesian estimation of CAR models, which accounts for uncertainty in the data, also helps smooth estimates over space so that areas with few neighbors are borrowing information from other regions (LeSage and Pace 2004).

68.8 Conclusion

The multilevel statistical approach – an approach that explicitly models the correlated nature of the data arising either due to sampling design or because populations are clustered – is a powerful tool for ecological inference.

From a substantive perspective, these multilevel methods circumvent the problems associated with **ecological fallacy** (the invalid transfer of results observed at the ecological level to the individual level), individualistic fallacy (occurs by failing to take into account the ecology or context within which individual relationships happen),

and atomistic fallacy (arises when associations between individual variables are used to make inferences on the association between the analogous variables at the group/ecological level). The issue common to the above fallacies is the failure to recognize the existence of unique relationships being observable at multiple levels and each being important in its own right. Specifically, one can think of an individual relationship (e.g., individuals who are poor are more likely to have poor health), an ecological/contextual relationship (e.g., places with a high proportion of poor individuals are more likely to have higher rates of poor health), and an individual-contextual relationship (e.g., the greatest likelihood of being in poor health is found for poor individuals in places with a high proportion of poor people). Multilevel models explicitly recognize the level-contingent nature of relationships.

Multilevel approaches also enable researchers to obtain statistically efficient estimates of fixed regression coefficients. Multilevel models provide correct standard errors and thereby robust confidence intervals and significance tests. These generally will be more conservative than the traditional ones that are obtained simply by ignoring the presence of clustering. More broadly, multilevel models allow a more appropriate and realistic specification of complex variance structures at each level. Multilevel models are also precision weighted and capitalize on the advantages that accrue as a result of “pooling” information from all the neighborhoods to make inferences about specific neighborhoods.

Developing and interpreting multilevel applications for inferences at the ecologic level entail several considerations. First, subject matter expertise and theory should drive choice of higher levels in a multilevel analysis. Second, establishing the relative importance of context and composition is probably more apparent than real, and necessary caution must be exercised while conceptualizing and interpreting the compositional and contextual sources of variation. Disentangling context from composition requires exposure and covariate data be collected at multiple levels, and that higher-level variables are not completely confounded with lower-level attributes. Third, it is important that the sample of neighborhoods belong to well-defined population of neighborhoods such that the sample shares exchangeable properties that are essential for robust inferences. In other words, units that are treated in statistical models as though they come from the same distribution should actually share commonalities that justify this treatment. Fourth, it is important to ensure adequate sample size at all levels of analysis. In general, if the research focus is essentially on neighborhoods, then clearly the analysis requires more neighborhoods (as compared to more individuals within a neighborhood). Lastly, like all quantitative procedures, the ability of multilevel models to make **causal inferences** is limited, and innovative strategies including randomized neighborhood-level research designs (via trials or natural experiments) in combination with multilevel analytical strategy may be required to convincingly demonstrate causal effects of social contexts such as neighborhoods. However, to the extent that experimental studies of contextual effects are few and far between, applying multilevel frameworks to observational data is a key aspect of neighborhood research.

References

- Aitkin M, Anderson D, Hinde J (1981) Statistical modelling of data on teaching styles (with discussion). *J R Stat Soc* 144:148–161
- Apparicio P, Abdelmajid M, Riva M, Shearmur R (2008) Comparing alternative approaches to measuring the geographical accessibility of urban health services: distance types and aggregation-error issues. *Int J Health Geogr* 7(1):7
- Bennett N (1976) Teaching styles and pupil progress. Open Books, London
- Berkman LF, Kawachi I (2000) Social epidemiology. Oxford University Press, New York
- Best N, Spiegelhalter D, Thomas A, Brayne C (1996) Bayesian analysis of realistically complex models. *J R Stat Soc A Stat Soc* 159:323–342
- Blakely T, Subramanian SV (2006) Multilevel studies. In: Oakes JM, Kaufman JS (eds) Methods in social epidemiology, vol 7. Jossey-Bass, San Francisco
- Blakely TA, Woodward AJ (2000) Ecological effects in multi-level studies. *J Epidemiol Community Health* 54(5):367–374
- Braumoeller BF (2006) Explaining variance; or, stuck in a moment we can't get out of. *Polit Anal* 14(3):268–290. doi:10.1093/pan/mpj009
- Browne WJ, Subramanian SV, Jones K, Goldstein H (2005) Variance partitioning in multilevel logistic models that exhibit overdispersion. *J Roy Stat Soc: Ser A (Stat Soc)* 168(3):599–613. doi:10.1111/j.1467-985X.2004.00365.x
- Diez Roux AV (2002) A glossary for multilevel analysis. *J Epidemiol Community Health* 56(8):588–594
- Fischer MM, Getis A (2010) Handbook of applied spatial analysis: software tools, methods and applications. Springer, Berlin
- Goldstein H (2003) Multilevel statistical models. Edward Arnold, London
- Goldstein H, Browne W, Rasbash J (2002) Partitioning variation in multilevel models. *Underst Stat* 1(4):223–231. doi:10.1207/S15328031US0104_02
- Jones K, Bullen N (1994) Contextual models of urban house prices: a comparison of fixed-and random-coefficient models developed by expansion. *Econ Geogr* 70:252–272
- Jones K, Moon G (1993) Medical geography: taking space seriously. *Prog Hum Geogr* 17(4):515–524
- Katz LF, Kling JR, Liebman JB (2000) Moving to opportunity in Boston: early results of a randomized mobility experiment. National Bureau of Economic Research, Cambridge, MA
- Kawachi I, Berkman LF (2003) Neighborhoods and health. Oxford University Press, Oxford
- Krieger N, Chen JT, Waterman PD, Rehkopf DH, Subramanian SV (2005) Painting a truer picture of US socioeconomic and racial/ethnic health inequalities: the Public Health Disparities Geocoding Project. *Am J Public Health* 95(2):312–323. doi:10.2105/AJPH.2003.032482
- Larsen K, Merlo J (2005) Appropriate assessment of neighborhood effects on individual health: integrating random and fixed effects in multilevel logistic regression. *Am J Epidemiol* 161(1):81–88. doi:10.1093/aje/kwi017
- LeSage JP, Pace RK (2004) Spatial and spatiotemporal econometrics. Emerald Group Publishing, Amsterdam
- Ludwig J, Sanbonmatsu L, Gennetian L, Adam E, Duncan GJ, Katz LF, Kessler RC et al (2011) Neighborhoods, obesity, and diabetes—a randomized social experiment. *N Engl J Med* 365(16):1509–1519. doi:10.1056/NEJMsa1103216
- Merlo J, Asplund K, Lynch J, Råstam L, Dobson A, World Health Organization MONICA Project (2004) Population effects on individual systolic blood pressure: a multilevel analysis of the World Health Organization MONICA project. *Am J Epidemiol* 159(12):1168–1179. doi:10.1093/aje/kwh160
- Moon G, Subramanian S, Jones K, Duncan C, Twigg L (2005) Area-based studies and the evaluation of multilevel influences on health outcomes. In: Bowling A, Ebrahim S (eds) Handbook of health research methods: investigation, measurement and analysis. Open University Press, Berkshire, pp 266–292

- Oakes JM (2004) The (mis)estimation of neighborhood effects: causal inference for a practicable social epidemiology. *Soc Sci Med* 58(10):1929–1952. doi:10.1016/j.socscimed.2003.08.004.
- Raudenbush SW, Bryk AS (2002) Hierarchical linear models: applications and data analysis methods, vol 1. Sage Publications, Thousand Oaks
- Rosenbaum JE (1995) Changing the geography of opportunity by expanding residential choice: lessons from the Gautreaux program. *Hous Policy Debate* 6(1):231–269
- Roux AVD (2008) Next steps in understanding the multilevel determinants of health. *J Epidemiol Community Health* 62(11):957–959. doi:10.1136/jech.2007.064311
- Sampson RJ, Morenoff JD, Gannon-Rowley T (2002) Assessing “neighborhood effects”: social processes and new directions in research. *Annu Rev Sociol* 28:443–478
- Subramanian SV (2004a) Multilevel methods, theory and analysis. In: Encyclopedia of health and behavior, vol 2. Sage Publications, Thousand Oaks, pp 602–608
- Subramanian SV (2004b) The relevance of multilevel statistical methods for identifying causal neighborhood effects. *Soc Sci Med* 58(10):1961–1967. doi:10.1016/S0277-9536(03)00415-5
- Subramanian SV, Glymour MM, Kawachi I (2007) Identifying causal ecologic effects on health: a methodological assessment. In: Macrosocial determinants of population health. Springer Media, New York, pp 301–331, Retrieved from <http://www.springerlink.com.ezp-prod1.hul.harvard.edu/content/p3777604031n2513/>
- Subramanian SV, O’Malley AJ (2010) Modeling neighborhood effects: the futility of comparing mixed and marginal approaches. *Epidemiol (Cambridge, MA)* 21(4):475–478. doi:10.1097/EDE.0b013e3181d74a71, discussion 479–481
- Subramanian SV, Jones K, Kaddour A, Krieger N (2009) Revisiting Robinson: the perils of individualistic and ecologic fallacy. *Int J Epidemiol* 38(2):342–360. doi:10.1093/ije/dyn359, author reply 370–373

Sergio J. Rey

Contents

69.1	Introduction	1365
69.2	Spatial Dynamics in Regional Science	1366
69.3	Exploratory Space-Time Data Analysis	1372
69.3.1	ETSDA Methods	1373
69.3.2	ESTDA Methods	1377
69.4	Conclusion	1381
	References	1382

Abstract

This chapter provides an overview of spatial dynamics in the field of regional science. After defining the context of spatial dynamics and the alternative conceptualizations of space and time, the chapter surveys the various areas of substantive interest where spatial dynamics come to the fore. A second focus is on the methodological and technical issues surrounding the methods of space-time data analysis. Here the emphasis is on exploratory methods for space-time data focusing on the evolution of spatial patterns as well as the identification of temporal dynamics that cluster in space.

S.J. Rey

GeoDa Center for Geospatial Analysis and Computation, School of Geographical Sciences
and Urban Planning, Arizona State University, Tucson, AZ, USA
e-mail: srey@asu.edu

69.1 Introduction

All human activity happens somewhere in space and time. As a consequence, spatial dynamics and space-time data analysis are of interest across a wide array of cognizant fields. Regional science is no exception as the consideration of the spatial and temporal domains in the theoretical and empirical analysis of socioeconomic phenomena is now a central theme in the discipline.

There are a number of forces that have given rise to this prominence. The first is technological in nature and reflects the increasing availability of longitudinal spatial data sets which have been made possible by the rise in geospatial technologies such as Global Positioning Services (GPS), network sensors, and areal photography.

The second development is theoretical in nature and reflects a shift in the focus of substantive theory from an initial view of independent agents operating in a spaceless world to one where geographical space becomes a key dimension of the underlying theory. The rise of spatial economics and the new economic geography are emblematic of this phase, which perhaps was prematurely labeled as “space: the final frontier” (Krugman 1998).

More recently, however, the frontier has been expanded by the recognition that in order to posit causal relationships, both the “where” question and the “when” question need to be addressed (Cressie and Wikle 2011). In this sense space-time is seen as the next frontier. Similarly, Goodchild (2008) sees spatiotemporal concerns as one of the key challenges facing the development of future GIS research. As regional science has increasingly adopted methods from geographical information science, it too faces these newly emerging challenges.

This chapter provides an overview of spatial dynamics in the field of regional science. After defining the context of spatial dynamics and the alternative conceptualizations of space and time, the chapter surveys the various areas of substantive interest where spatial dynamics come to the fore. A second focus is on the methodological and technical issues surrounding the methods of space-time data analysis. Here the emphasis is on exploratory methods for space-time data focusing on the evolution of spatial patterns as well as the identification of temporal dynamics that cluster in space.

69.2 Spatial Dynamics in Regional Science

To situate the discussion, it is important to first consider the different ways that space and time, as well as spatial dynamics, have been conceptualized in practice. Both space and time have been viewed in a variety of ways, which makes the consideration of space-time even more complex.

From a geographical perspective, two different conceptualizations of space have been used: the object and field views. In the former, the world is seen as populated by discrete homogeneous units such as factories, homes, roads, lakes, and rivers that are located using some form of geographical coordinate system. Often these are represented using a vector data model that relies on points, lines, and polygons.

Alternatively, in a field-based ontology, the focus is on the spatial variation in some phenomenon (such as temperature, risk exposure, or elevation) across space that can conceptually be observed at any location and is thus spatially continuous. For fields, a raster data model is adopted where pixels, grid cells, or voxels are used to exhaust space. Conventionally, the object view has been dominant in regional science given the focus on macro-spatial economic, demographic, and social phenomena.

In a similar way time has been treated as both discrete and continuous in regional science. By and large the former perspective is more commonly encountered, and this is in large part due to practical considerations related to the way data series are recorded. At the same time, work on continuous-time models has remained largely theoretical with emphases on understanding the dynamic trajectories following from various forms of growth and interaction specifications.

Consideration of the discrete versus continuous categorization of space and time gives rise to a four-way classification. By far the most commonly employed classification in regional science is the discrete-space, discrete-time classification, where, for example, space is organized as areal units (states, provinces, counties) and time is measured on an monthly, quarterly, annual, or decadal frequency. Although this is the dominant approach, exceptions can be found. The work by Arbia and Paelinck (2003) on regional convergence is an example of a discrete-space, continuous-time approach. Conversely, in Duranton and Overman (2008) the focus is on the distribution of individual firms over a continuous space but in discrete time. Finally, the theoretical work of Fujita et al. (2001) specifies models for optimizing agents in continuous time and continuous space.

The intersection of space-time provides a mechanism to move beyond the traditional cross-sectional focus that has long dominated empirical work in regional science. The estimation of econometric relationships using cross-sectional data rests on the assumption that the underlying process has reached a state of equilibrium. At the same time, however, many of the phenomena of interest to regional science are often viewed from an adjustment or disequilibrium perspective. Methodologically, the latter requires space-time data in order for empirical investigation to be possible.

Along with the discrete versus continuous view of space and time, existing work in regional science can also be characterized according to the relative dominance of one of these dimensions. For example, work in spatial econometrics and exploratory spatial data analysis has been predominately concerned with the analysis of spatially referenced data at one point in time. Here the spatial dimension is central, while the temporal dimension is ignored or radically reduced to $t = 1$. These could be referred to as large n , extremely small t type studies. In contrast, there are studies of single regions where the emphasis is on the dynamic behavior of the individual regional economic or demographic system as in the case of early work on single-region macroeconometric models (Bolton 1985). These are extremely small $n = 1$, large t -type studies where the time series dimension is exploited to parameterize models for a small number of regions.

More data rich studies arise when multiple regions and time periods are analyzed. Prominent examples of large n small t type studies would include most β -convergence studies where the growth rate over two points of time is analyzed for a large set of n regions (Rey and Le Gallo 2009). Further distinctions must be drawn between studies that involve multiregional versus interregional analysis. In the former multiple regions (i.e., $n > 1$) provide an increased sample size for model parameterization, but the regions themselves may not interact. By contrast, interregional models explicitly incorporate interactions between the regions. This provides for further differentiation between dynamic processes and spatial processes. A dynamic process is one that transitions over time, for example, in the study of regional business cycles, a focus on the characteristic of a particular region's cycle behavior. A spatial process is distinct from a temporal process in that the former does not act on a single location but involves interaction across different locations that transpire over time.

Spatial dynamics pertains to a dynamic process that is spatially dependent (Irwin 2010). Spatial dynamics are relevant to many areas of applied and theoretical regional science. A prominent substantive motivation for spatial dynamics is in the study of optimal currency areas (OCA). One of the key criteria for a group of economies to be a candidate for an OCA is that their business cycles display a high degree of co-movement or synchronization so that a single monetary policy could be effectively employed (Partridge and Rickman 2005). An additional example is the related literature that attempts to identify the leading-lagging relationships between pairs of regional economies using various Granger causality and vector autoregressive models (LeSage and Reed 1990).

A further distinction arises from a consideration of two related, but distinct, concepts: comparative statics versus spatial difference. In comparative statics, a system is compared at two, or more, points in time to identify shifts or changes in the state of the system in discrete time, such as movements of, or along, a supply or demand curve. In a geographical context, the analogy is one of spatial difference – that is, comparing the articulation of a process at two or more different locations, but at the same moment in time.

In practice the researcher is faced with the similar challenge of trying to make inferences about the process that may be responsible for the temporal change or spatial differences observed. Two broad strategies have been adopted here. The first relies on so-called pattern models which can be viewed as analogous to a reduced form model in that they focus on describing the evolution of patterns which reflect the operation of some underlying process. Alternatively process-based models are akin to a structural form in that model parameters are tied directly to behavioral units in the underlying substantive theory for the process under study. A key challenge to linking substantive theory to space-time patterns is that substantive theories are often not detailed enough to make this linkage.

Not only do the different conceptualizations result in different representations of phenomena in space and time, but they tend to be more prevalent in certain types of space-time domains and can also require different analytical and statistical methods as is explored below.

In addition to these different conceptual frameworks for understanding space-time data analysis, one can also approach the topic from the different space-time domains that appear in substantive studies. Goodchild (2008) has offered a taxonomy of space-time domains that considers five different areas of inquiry.

Tracking the movement of individuals within a city using GPS devices provides a new way to understand human activity patterns within an urban context. These can be seen as modern extensions of foundational work of Hägerstrand (1970) formalization of tracking individual activity spaces as space-time prisms. In the current implementation, the masses of data generated from real-time network sensors, RFID, GPS, and social media postings (i.e., Twitter) have generated an active literature developing interesting new ways to analyze such data, and these methods are driving new innovations in transportation planning.

The second domain for space-time analysis concerns change detection or so-called snapshots. Time series of remotely sensed images of urban areas (Yang et al. 2003) can be used to analyze changes in urban morphology as well as trends in rural–urban land-use conversion that are becoming increasingly important to the understanding of coupled human and natural systems. Formal modeling of the evolution of such spatial patterns has been a central concern in health and environmental applications (Abellán et al. 2008; Wikle and Cressie 2000).

Polygon coverage, the third space-time domain, focuses on changes in attribute values for areal units over time. As we return to below in Sect. 69.3, a rich set of methods has been suggested to characterize these space-time dynamics. An important challenge in the polygon coverage domain arises when the reporting units and boundaries as well as the attribute values change over time. We return to this challenge below.

The fourth space-time domain shifts the focus to the raster data model and employs cellular automata (CA) models in which a set of states for each raster cell are specified together with a set of rules that determine the state transitions through time. Emblematic of this line of inquiry is the work on urban development (Clarke et al. 2007). Closely related to CA models are agent-based models in which space is viewed as populated with discrete agents, which could be either geographical features or actors, that are embodied with rules governing their behavior.

The final space-time domain concerns events and transitions. The classic example is Minard's map in Fig. 69.1 depicting Napoleon's march and retreat on Russia. This visualization combines spatial and temporal dimensions along with the depiction of temperature and information on the size of Napoleon's army in a highly complex representation. Although to date it has not been done, there is no reason why such methods could not be applied to events in regional science, such as business cycles or interstate migration patterns. Currently the business cycles are studied at two spatial scales, with attention at one level on the individual business cycles of states and how those cycles may be correlated or synchronized with the cycles of other states, or how they may be related to the cycle at a higher

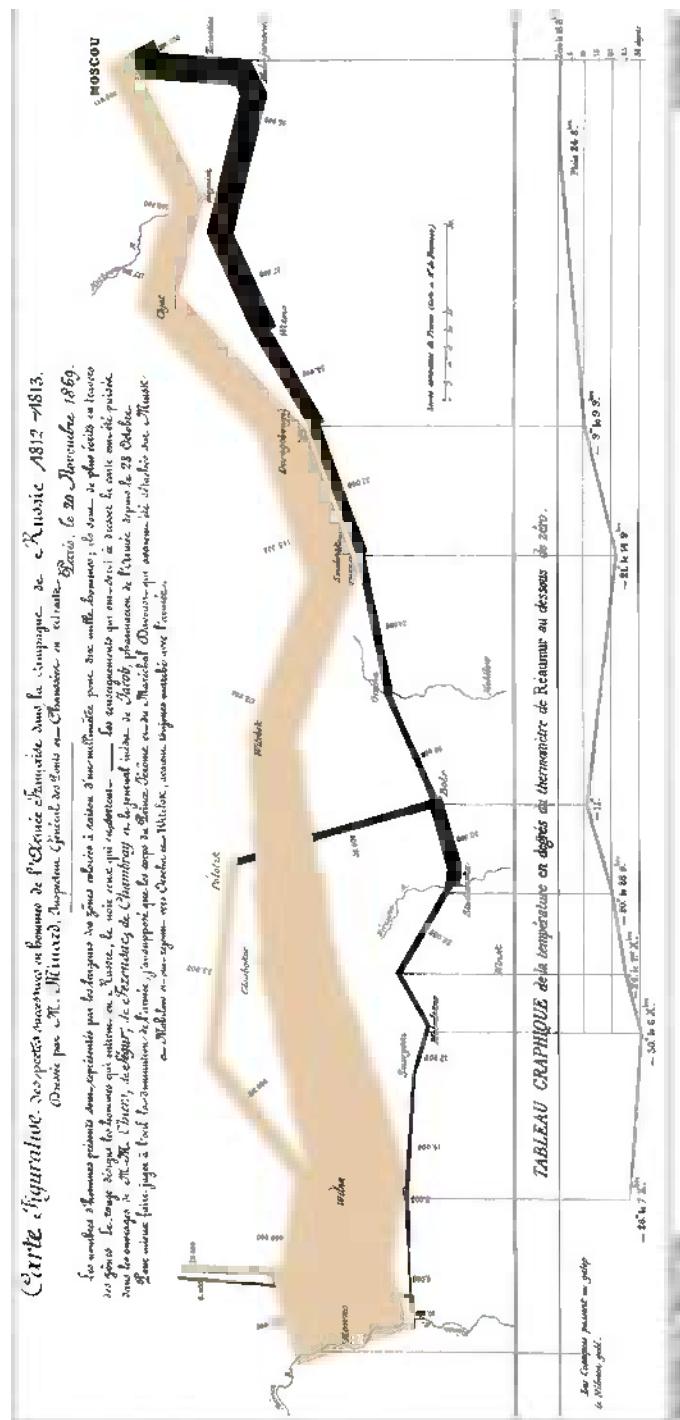


Fig. 69.1 Minard's map

spatial scale, say nation. The event framework provides the potential for integrating these seemingly distinct cycles into a unified cycle that may be articulated across space at different points and time.

The empirical analysis of each of these space-time domains relies on data that is organized in some space-time framework. Spatial data has conventionally been organized along the following taxonomy. Point data are used to represent the locations of individual events and the interest rests on the resulting spatial pattern of those locations together with any additional attribute data about the events – as is the case for marked point patterns used in the study of firm location and retail competition.

Geostatistical data arises when observations at fixed locations are obtained on some spatial phenomena that conceptually varies continuously over some spatial domain. Unlike point patterns, where the interest is on the pattern of the locations, in geostatistical data the focus is on the variation in the attribute across the fixed observation (or sample points), and models of this variation are used to develop predictions of the phenomena at target sites.

Arguably the most commonly encountered type of spatial data in regional science is lattice or areal unit data. Here space is viewed as partitioned into discrete areal units, and variables are measured for each of the units. The focus is on understanding the variation in the attribute across spatial units; however, unlike the case of geostatistical data, interpolation or spatial prediction is meaningless since the areal units exhaust the spatial domain.

Although these three types of spatial data form the core of the taxonomy in most spatial statistics texts, there are other types of spatial data that are commonly encountered in regional science research. Chief among these is network data which is prominent in many transportation studies. Network data are also encountered in various optimization models, and network concepts play a central role in defining spatial relationships between areal units in the analysis of lattice data. Networks are increasingly being used to model social interactions as in the growing literature on social networks with interest in embedding these social networks in geographic space.

Space-time data opens up a number of important ways to address fundamental problems that confront researchers working in either a cross-sectional or time series framework. In spatial analysis there has long been tension between so-called complete spatial heterogeneity where each location can be seen as unique, and more general lawlike constructions that apply in all locations. From a data analysis perspective, the former is a nonstarter since insufficient degrees of freedom are available – in a sense the number of parameters grows with the number of areas under consideration since each place is unique and requires its own parameter values. Enforcing spatial homogeneity is one way to reduce the parameter space and make inference tractable. This comes at a cost of course of imposing uniformity on the processes over space.

With space-time data there is more flexibility in the types of specifications that could be considered. In cases where a long enough time series is available, the formally intractable problem of allowing each place to have its own model

as it were can now be relaxed. Indeed, certain models adopted in practice actually require long time series for their use, as in the case of the Hildreth Prescott filter used to study regional business cycle behavior. Use of the HP filter with shorter series is known to introduce distortions (Partridge and Rickman 2005).

While the temporal dimension allows for a relaxation of the spatial homogeneity assumption, treatment of spatial dependence in a dynamic context must also be considered. Rather curiously, most approaches that consider spatial dynamics assume that the form of the interaction process is stable over time. In other words the strength of the spatial dependence is often held constant. For example, work on the identification of leading and lagging regions employs Granger causality type frameworks that exhaust the temporal dimensions to estimate the nature of the dynamic relationships between each pair of economies. The identified temporal lags are then assumed to hold over the entire time series.

While this approach does allow for spatial variation in the degree of spatial dependence (since each pair of economies can have distinct lead-lag relationships), it comes at a cost of assuming the spatial dynamics are temporally invariant. Such an assumption may be overly restrictive, since research in the area of regional income convergence (Rey and Le Gallo 2009) and business cycles (Partridge and Rickman 2005) is suggesting that the strength of spatial interaction in regional macro-series is often not constant over time.

Research that is extending these different types of spatial data supports and the associated analytical methods to include a temporal dimension is only at very embryonic stage of development. Most of the work on developing analytical methods for space-time data in regional science has focused on areal unit or polygon data and has adopted exploratory focus. An overview of the key directions in this regard is provided in the remainder of the chapter.

69.3 Exploratory Space-Time Data Analysis

Methods for exploratory space-time data analysis for lattice data can be organized in a number of ways. The first is to make a distinction between those that have their origins as cross-sectional methods that have been extended to incorporate a temporal dimension. Alongside of these are methods that were originally temporal exploratory data analysis (EDA) methods that were modified to incorporate space. The former group of methods can be viewed as studying the evolution of spatial patterns in time, while the latter switches the perspective to put temporal dynamics into space. In other words, the first group of methods views the spatial dimension from a temporal perspective, while in the latter the spatial signature of dynamic patterns becomes the focus.

To distinguish between these two sets of methods in what follows the acronym, ETSDA is used for the approaches that have their origins in the temporal domain but have been extended to incorporate space, while ESTDA is used for the originally spatial methods that have been extended to incorporate time. Although the perspectives are distinct across these two groups, in both cases there are

methods that are numerical and sometimes coupled with novel visualization methods which are also discussed.

69.3.1 ETSDA Methods

A main branch of the ETSDA literature begins with discrete Markov chains. A Markov chain is a particular type of dynamic stochastic process $\{X(t)|t \in T\}$ that satisfies the following condition. For any $t_0 < t_1 < \dots < t_n$,

$$\begin{aligned} P[X(t_n) = x_n | X(t_{n-1}) = x_{n-1}, X(t_{n-2}) = x_{n-2}, \dots, X(t_0) = x_0] \\ = P[X(t) = x_n | X(t_{n-1}) = x_{n-1}] \end{aligned} \quad (69.1)$$

This condition implies that the conditional distribution function of $X(t_n)$ only depends on $X(t_{n-1})$. In other words, given the present state of the process, the future state of the process is independent of the past.

A discrete-state Markov process is one in which the random variable X takes on one of n unique values. Such a Markov process is known as a Markov chain in which case Eq. (69.1) takes the following form:

$$P[X_k = j | X_{k-1} = i, X_{k-2} = n, \dots, X_0 = m] = P[X_k = j | X_{k-1} = i] = p_{i,j,k} \quad (69.2)$$

where $p_{i,j,k}$ is the state transition probability reflecting the conditional probability that the process will be in state j at time k given that it is in state i at time $k-1$. For a time homogeneous Markov chain, the transition probabilities are time invariant, which implies the following:

$$P[X_k = j | X_{k-1} = i, X_{k-2} = n, \dots, X_0 = m] = P[X_k = j | X_{k-1} = i] = p_{i,j} \quad (69.3)$$

These transition probabilities satisfy the following conditions:

1. $0 \leq p_{i,j} \leq 1$
2. $\sum_j p_{i,j} = 1, \forall i, i = 1, 2, \dots, n$.

Given the n states, the transition probability matrix is

$$P = \begin{bmatrix} p_{1,1} & p_{1,2} & \dots & p_{1,n} \\ p_{2,1} & p_{2,2} & \dots & p_{2,n} \\ \vdots & \vdots & \vdots & \vdots \\ p_{n,1} & p_{n,2} & \dots & p_{n,n} \end{bmatrix} \quad (69.4)$$

Estimation of the transition probabilities can be based on maximum likelihood assuming time homogeneity:

$$\hat{p}_{i,j} = \frac{v_{i,j}}{\sum_j v_{i,j}} \quad (69.5)$$

where $v_{i,j}$ is the number of observed chain transitions from state i to state j .

Markov chains have played a central role in the literature on regional income convergence, following the pioneering work by Quah (1993). The typical approach is to discretize the distribution of per capita incomes or gross regional product measured over n regions into k classes in each time period, giving the discrete distribution π_t . Next, transition probabilities across each of the these k classes of this distribution are formalized in a $k \times k$ matrix of transition probabilities:

$$P = \begin{bmatrix} p_{1,1} & p_{1,2} & \cdots & p_{1,k} \\ p_{2,1} & p_{2,2} & \cdots & p_{2,k} \\ \vdots & \vdots & \vdots & \vdots \\ p_{k,1} & p_{k,2} & \cdots & p_{k,k} \end{bmatrix} \quad (69.6)$$

where $p_{i,j}$ is the probability of an economy moving from class i of the distribution at time t into class j in period $t + 1$. Two key assumptions are often relied upon in regional convergence applications of Markov chains. The first is that temporal homogeneity holds

$$P_t = P_{t+1} = \dots = P_{T-1} = P_T \quad (69.7)$$

The second assumption is that the chain is first order:

$$P(X_t = j | X_{n-1} = k, \dots, X_0 = l) = P(X_t = j | X_{n-1} = k) \quad (69.8)$$

This means that the only relevant information is the state of the chain in the preceding period – the state of the chain from more distant periods has no effect on future dynamics.

These two assumptions allow a mapping of the distribution between any pair of periods:

$$\pi'_{t+1} = \pi'_t P \quad (69.9)$$

or, more generally,

$$\pi'_{t+b} = \pi'_t P^b \quad (69.10)$$

A final assumption that is also sometimes made is that the chain is irreducible. Formally, for each pair of states there is some length of time $v_{i,j}$ where

$$\pi_{i,j}^{v_{i,j}} > 0 \quad \forall i, j \quad (69.11)$$

meaning that movements between any pair (i, j) of states in the distribution are possible over some time horizon.

Homogeneity and irreducibility combined implies that the chain will have a steady-state distribution π_* where

$$\pi'_* = \pi'_* P \quad (69.12)$$

with

$$\Pi_* = P_{v \rightarrow \infty}^v \quad (69.13)$$

The rows of the matrix Π_* will be identical and represent the long-run or ergodic income distribution π_* .

In the convergence literature this framework has been used to study a number of issues, including the time required to achieve convergence, the extent of polarization, and the degree of intradistributional mobility (Rey and Le Gallo 2009). In addition to regional convergence, Markov chains have seen application in the area of city-size distributions (Black and Henderson 2003). It should also be noted that approaches such as the stochastic kernel which is the continuous equivalent of the transition probability matrix that overcomes some of the inherent shortcomings of a discrete-space setup have been suggested (Fischer and Stumpner 2008).

The classic Markov framework applied above has been extended to incorporate a spatial dimension in a number of ways. The first is through regional conditioning (Quah 1993) in which the distribution of neighbor-relative incomes is mapped into the distribution of nation-relative incomes, with the former obtained by normalizing incomes relative to the average of those of a region's geographical neighbors:

$$yr_{i,t} = \frac{y_{i,t}}{\sum_j^n w_{i,j} y_{j,t}} \quad (69.14)$$

where $y_{i,t}$ is income in region i in time period t and $w_{i,j}$ is an element of a row-standardized spatial weights matrix expressing the neighbor relation between each pair of economies. The national-relative distribution is defined using

$$yn_{i,t} = \frac{y_{i,t}}{1/n \sum_j^n y_{j,t}} \quad (69.15)$$

The regional conditioning allows for an analysis of the degree of spatial clustering in the regional income distribution since the two discrete relative distributions (69.14) and (69.15) should be independent if incomes were randomly distributed in space. This would be reflected in a diagonally dominant transition matrix that maps Eq. (69.14) into Eq. (69.15).

Spatial Markov: Regional conditioning, however, considers spatial autocorrelation at one point in time, so in a sense it is not a dynamic Markov chain. Rey (2001) extended the classic dynamic Markov chain to include a spatial component through the concept of a *spatial Markov chain*. Defining Markov chains conditioned on

different classes of the spatial lag (defined using the denominator of Eq. (69.14)) allows for an assessment of the role of spatial context in shaping the transitional dynamics. A growing body of research reveals contextual effects of a spatial nature as transition probabilities show clear dependence on the relative incomes of neighboring economies (Bosker 2009; Hammond 2004; Le Gallo 2004).

Spatial Rank Dynamics: A second subclass of ETSDA methods departs from the use of various bivariate correlation methods to explore dynamics. Borrowing from work on map comparisons where different types of correlation methods are applied to two contemporaneous map patterns (Lloyd and Steinke 1977), it was a short step to apply the same framework for maps from two different time periods (rather than for different variables at the same point in time).

Interestingly the methods used are classical, or spatial, correlation methods. More specifically, a traditional rank correlation statistic is applied:

$$\tau_{t,t-1} = \frac{C_{t,t-1} - D_{t,t-1}}{n(n-1)/2} \quad (69.16)$$

where $C_{t,t-1}$ is the number of concordant pairs of observations and $D_{t,t-1}$ is the number of discordant pairs between time periods $t-1$ and t . A pair of regions i,j is concordant if

$$(r_{i,t} - r_{j,t})(r_{i,t-1} - r_{j,t-1}) > 0 \quad (69.17)$$

where $r_{i,t}$ is the rank of region i in period t . If the sign of the rank difference product is negative, the pair of regions is discordant. A close inspection of this statistic reveals that the only position that matters here is the relative location of each area in the rank distribution. The geographical location of the observation is ignored.

Rey (2004) has suggested an extension of this traditional rank correlation measure to incorporate a spatial dimension. Using a spatial concordance decomposition,

$$\tau_{t,t-1} = \frac{CG_{t,t-1} + CN_{t,t-1} - DG_{t,t-1} - DN_{t,t-1}}{n(n-1)/2} \quad (69.18)$$

where the number of contiguous pairs is separated into those involving geographical neighbors (G) and those that are not neighbors (N):

$$C_{t,t-1} = CG_{t,t-1} + CN_{t,t-1} \quad (69.19)$$

and the same decomposition is used for the discordant pairs. This can be viewed in a number of ways. First, the contributions of the two types of pairs to the spatial level of concordance or discordance can be evaluated. Alternatively, the degree of rank concordance for the two sets of pairs of regions can be contrasted, by noting:

$$\tau_{t,t-1} = \omega_G \tau_{G,t,t-1} + \omega_N \tau_{N,t,t-1} \quad (69.20)$$

where ω_G is the share of all pairs that involve geographic neighbors and

$$\tau_{G,t,t-1} = \frac{CG_{t,t-1} - DG_{t,t-1}}{\omega_G n(n-1)/2} \quad (69.21)$$

This provides insight as to the role of spatial dependence in the overall degree of temporal rank concordance. By contrasting the degree of rank correlation for neighboring pairs of regions with that of geographically separated pairs, the degree to which distributional mixing is spatially clustered can now be estimated.

69.3.2 ESTDA Methods

For ESTDA methods the point of departure is a method that was originally developed for cross-sectional analysis. Typically this takes the form of a method designed to detect spatial autocorrelation, either as a global or local form. From here, a dynamic component is added to enable the analysis of spatial dynamics. A common strategy is the repeated application of Moran's I to a temporal sequence of measurements on a variable for regions. Moran's I in period t is given as

$$I_t = \frac{n}{S_0} \frac{\sum_{i=1}^n \sum_{j=1}^n z_{i,t} w_{i,j} z_{j,t}}{\sum_{i=1}^n z_{i,t}^2} \quad (69.22)$$

where $S_0 = \sum_{i=1}^n \sum_{j=1}^n w_{i,j}$ and $z_{i,t} = y_{i,t} - \bar{y}_{i,t}$, and $w_{i,j}$ is as defined in Eq. (69.14). Examples of this approach can be found in the convergence literature where a common finding has been that time series of global Moran's I values for per capita income/product display significant positive spatial autocorrelation over time but also the strength of that spatial clustering exhibited substantial temporal variation (Rey and Le Gallo 2009).

The same comparative static design has also been used to explore how local measures of spatial autocorrelation change over time. In a cross-sectional setting, local measures provide indications of departures of the overall pattern of global spatial dependence or allow for the detection of spatial outliers, hot spots and/or cold spots (Anselin 1995). In a similar vein, when extended to a space-time setting, this provides a useful complement to the comparative static analysis of global spatial autocorrelation dynamics. The focus remains on the relative stability of local spatial association patterns through time which is enabled through a comparison of a series of snapshots. The situation is more complex in the local case as now there are n values in each snapshot and there evolution over time increases the analytical demands relative to the global case in which only a single indicator is studied from a dynamic perspective.

Space-Time LISA: Closely related to the comparative static analysis of the LISA statistics is the bivariate LISA. The bivariate LISA modifies the original indicator

by shifting the time period for either the variable or the spatial lag of the variable. Two possibilities exist. The first consists of a temporal lag of the spatial lag:

$$L_{i,t} = z_{i,t} \sum_{j=1}^n w_{i,j} z_{j,t-k} \quad (69.23)$$

which relates the value at focal unit i in period t to that observed in its geographical neighborhood k periods previously. In the second form the shift is applied to the variate:

$$L_{i,t} = z_{i,t-k} \sum_{j=1}^n w_{i,j} z_{j,t} \quad (69.24)$$

The two forms lend themselves to different types of questions about local spatial dynamics that relate to the form and direction of the space-time spillover or diffusion. In the first form, if positive local space-time associate was indicated, this would be consistent with inward diffusion from the surrounding units into the core focal unit. By contrast, the temporal lag of the focal unit in the second form means that any positive association revealed would be consistent with diffusion originating from that unit and spreading outward to the neighbors.

The bivariate LISA moves the ESTDA methods from a comparative static view toward an explicit consideration of spatial dynamics in the sense that the dependence between a measurement at one location and point in time is being related to a different location at a different point in time. This is an important shift because it reduces the gap between the patterns being observed and the underlying dynamic process that may be responsible for that pattern.

Nowhere is this more apparent than in the distinction between apparent and true contagions. The former arises from a spatial pattern that could be consistent with a dynamic process such as the spread of an infectious disease through contact of individuals in close proximity to one another. A single map displaying spatial autocorrelation of disease incidence would be consistent with the operation of such a process. However, there are other processes that could also give rise to the same pattern – such as when the disease incidence may be driven by environmental factors (i.e., contaminated water supplies). Based on the single map, it is impossible to identify which is the operative process.

With maps from multiple time periods, however, the possibility to differentiate between true and apparent contagions now exists. The key signature difference would be for the map pattern to change over time in the case of true contagion reflecting the transmission over space, while the area of high incidence would remain spatially fixed in the case of apparent contagion – assuming the focal source was spatially immobile.

In the bivariate LISA, outward diffusion can be represented on a scatterplot where the x-axis has the rate in an initial period and the y-axis measures the spatial lag of the rate in the future period. For inward diffusion, the x-axis has the rate in the

future period, while the y-axis depicts the spatial lag in the previous period. In other words, the spatial lag is shifted either backward (inward diffusion) or forward (outward diffusion) in time to depict different forms of spatial dynamics.

There are several complications in association with this interpretation of the bivariate LISA as an indicator of spatial dynamics. One difficulty is that these patterns are also consistent with spatial dependence that is not changing over time. For example, if there was positive spatial autocorrelation that was constant over time, then a bivariate correlation of a variable at time t and its spatial lag at time $t \pm k$ are likely to be positive. Because the correlation is positive for both the forward and backward time-shift of the spatial lag, the approach would yield indications of both false inward and outward contagions, when in fact the underlying spatial dependence has been constant over time.

Directional LISA: A number of extensions to the LISA in a dynamic context have recently been suggested as ways to address these issues. The first is directional LISA that explicitly considers the movement of the LISA statistic between a pair of periods (Rey et al. 2011). More specifically, two Moran scatterplots are compared: one for the initial period (Fig. 69.2a) and one for the end time period (Fig. 69.2b). Based on these, the movement vectors are extracted to form the directional Moran scatterplot (Fig. 69.2c). The movement vectors can be either origin or destination standardized, which then permits a visualization of the direction, magnitude, and any biases in the spatial dynamics between the two periods (Fig. 69.2d).

The characteristics of these movement vectors can be summarized using several new visualization or inferential tools. For the former, a rose diagram depicts the relative frequency of movement vectors providing insights as to the concentration and potential biases of movements observed over a period (Fig. 69.2e). Coupled with this is a computationally based approach to inference in which the extent of spatial dependence in the movement vectors is tested against a null hypothesis that an observation and its spatial lag move independently over the time period.

LISA Markov: Closely related to both the directional LISA and the original space-time bivariate LISA is the LISA Markov (Rey 2001). This extends the focus to consider a sequence of moves by the local statistics, not just one period as is the case for the directional LISA. This relies on the quadrants of the Moran scatterplot which are now used to define the states for a discrete Markov chain. The four quadrants are I (H,H), II (L,H), III (L,L), and IV (H,L) with the first position indicating whether the observations are above or below the mean, while the second does the same but for the spatial lag. These four states give rise to 16 types of transitions.

The 16 transition types offer a rich taxonomy for characterizing spatial dynamics. For example, the issue of outward and inward diffusion that was encountered in the discussion of the bivariate space-time LISA can now be associated with particular moves in this taxonomy. Outward diffusion would be reflected in transitions where the spatial lag increases in value over time and the core either declines or remains high: (H,L)–(H,H) or (H,L)–(L, H). The two cases allow for a differentiation between saturation diffusion, in the former, and displacement diffusion in the latter. For inward diffusion the relevant moves

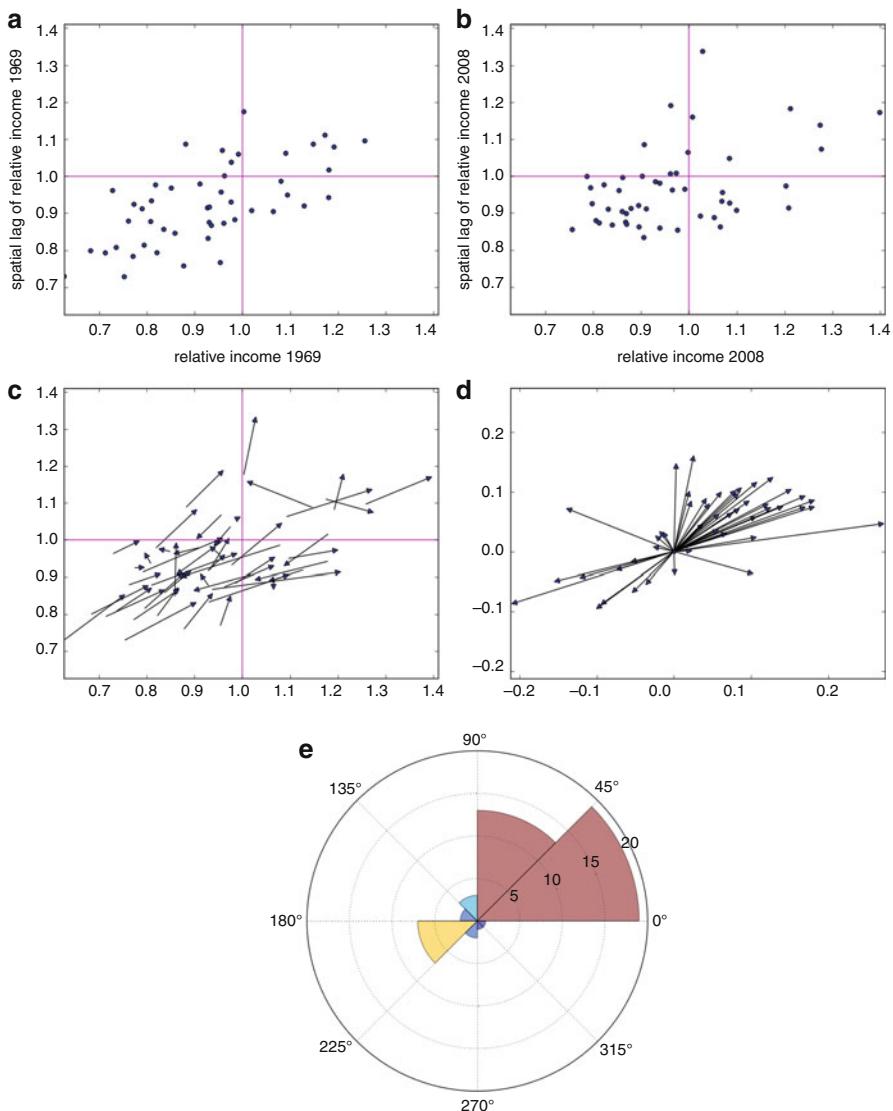


Fig. 69.2 Directional Moran scatter plots. (a) Moran scatter 1969. (b) Moran scatter 2008. (c) Unstandardized movement vectors. (d) Origin standardized movement vectors. (e) Rose diagram

would be (L,H)–(H,H) or (L,H)–(H,L); in either case the core increases over time, while the lag declines (displacement) or remains high (saturation).

Formal inference on these spatial dynamics has been suggested by Rey et al. (2012). The notion of a joint spatial Markov chain decomposes the spatial dynamics into two separate discrete chains, one for the original attribute and one for the spatial lag of this attribute. Each of these individual chains can occupy one of two

states in a given period, either (H) or (L). Letting $P(Y)$ represent the transition probability matrix for the original attribute chain and $P(WY)$ the transition probability matrix for the spatial lag of this attribute, under a null of independence (or lack of spatial dynamics), we have

$$P(\widetilde{Y}, \widetilde{WY}) = P(\widetilde{Y}) \otimes P(\widetilde{WY}) \quad (69.25)$$

where \otimes is the Kronecker product operator.

The estimated joint transition probability matrix from Eq. (69.25) is then compared to the observed joint transition probability matrix, $P(\widetilde{Y}, \widetilde{WY})$, and a formal test of the equality of these two transition matrices can be based on large sample theory for discrete Markov chains. Rejection of the equality hypothesis means that the two chains are non-separable. In other words, the dynamic transitions of the attribute values at a given location are not independent of the transitions of the spatial lag of these values.

In addition to providing a global test of spatial dynamics, comparison of the two estimated joint transition probability matrices allows for an identification of what types of moves are over, or under, represented in the observed spatial transitions, relative to the case where the dynamics displayed spatial randomness.

69.4 Conclusion

Regional science has long considered spatial dynamics as an organizing framework from which to view different regional phenomena. Regional growth theory by definition would not exist without a space-time framing. The inverted-U pattern proposed by Williamson (1965) of regional inequality provides a specific example where the level of regional inequality is viewed through a dynamic lens. While regional growth is a process that operates over space and time, the inverted-U framework is largely a-spatial as the regions are simply observational units used to measure dispersion in incomes. The actual location of these regions and issues of spatial interactions have not given explicit empirical treatment in this framework. As Miller (2006) has argued in the context of other areas of regional science, the spatial and temporal dimensions underlying human activity cannot be meaningfully separated. By the same token, regional science cannot be separated from a space-time framework or a consideration of spatial dynamics.

With recent technical and methodological developments in the areas of space-time data analysis, the possibility now exists to extend the traditional framework to include a richer spatial dynamics component, one that allows for a tighter linkage between abstract theoretical constructs and their empirical implementation. There are also gains to be had from applying some of these new measures of space-time dynamics to summarize outcomes of other types of modeling frameworks. For example, more comprehensive summaries of the predictions from land-use change models become possible. Similarly, the growing

use of agent-based models and cellular automata creates a need for efficient methods that can capture and summarize the spatial dynamics of these complex patterns generated from these frameworks.

Acknowledgments This research was funded in part by Award No. 2009-SQ-B9-K101 from the National Institute of Justice, Office of Justice Programs, US Department of Justice.

References

- Abellan J, Richardson S, Best N (2008) Use of space–time models to investigate the stability of patterns of disease. *Environ Health Perspect* 116(8):1111–1119
- Anselin L (1995) Local indicators of spatial association-LISA. *Geogr Anal* 27(2):93–115
- Arbia G, Paelinck J (2003) Spatial econometric modelling of regional convergence in continuous time. *Int Reg Sci Rev* 26:342–362
- Black D, Henderson V (2003) Urban evolution in the USA. *J Econ Geogr* 3(4):343–372
- Bolton R (1985) Regional econometric models. *J Reg Sci* 25(4):495–520
- Bosker M (2009) The spatial evolution of regional GDP disparities in the old and the new Europe. *Pap Reg Sci* 88(1):3–27
- Clarke K, Gazulis N, Dietzel C, Goldstein N (2007) A decade of SLEUTHing: lessons learned from applications of a cellular automaton land use change model. In Fisher P (ed) *Classics in IJGIS: twenty years of the International Journal of Geographical Information Science and Systems*. CRC Press, Boca Raton, pp 413–427
- Cressie N, Wikle C (2011) *Statistics for spatio-temporal data*. Wiley, New York
- Duranton G, Overman H (2008) Exploring the detailed location patterns of UK manufacturing industries using microgeographic data. *J Reg Sci* 48(1):213–243
- Fischer MM, Stumpner P (2008) Income distribution dynamics and cross-region convergence in Europe. *J Geogr Syst* 10(2):109–139
- Fujita M, Krugman P, Venables AJ (2001) *The spatial economy: cities, regions, and international trade*. MIT Press, Cambridge
- Goodchild M (2008) Combining space and time: new potential for temporal GIS. In: Knowles A (ed) *Placing history: how maps, spatial data and GIS are changing historical scholarship*. ESRI, Redlands, pp 179–198
- Hägerstrand T (1970) What about people in regional science? *Pap Reg Sci* 24(1):6–21
- Hammond GW (2004) Metropolitan/non-metropolitan divergence: a spatial Markov chains approach. *Pap Reg Sci* 83(3):543–563
- Irwin E (2010) New directions for urban economic models of land use change: incorporating spatial dynamics and heterogeneity. *J Reg Sci* 50(1):65–91
- Krugman P (1998) Space: the final frontier. *J Econ Perspect* 12(2):161–174
- Le Gallo J (2004) Space-time analysis of GDP disparities across European Regions: a Markov chains approach. *Int Reg Sci Rev* 27(2):138–163
- LeSage J, Reed J (1990) Testing criteria for determining leading regions in wage transmission models. *J Reg Sci* 30(1):37–50
- Lloyd R, Steinke T (1977) Visual and statistical comparison of choropleth maps. *Ann Assoc Am Geogr* 67(3):429–436
- Miller H (2006) Social exclusion in space and time. In: Axhausen K (ed) *Moving through nets: The physical and social dimensions of travel*. Elsevier, Amsterdam, pp 353–380
- Partridge M, Rickman D (2005) Regional cyclical asymmetries in an optimal currency area: an analysis using US state data. *Oxf Econ Pap* 57(3):373–397
- Quah DT (1993) Empirical cross-section dynamics in economic growth. *Eur Econ Rev* 37(2–3):426–434
- Rey SJ (2001) Spatial empirics for economic growth and convergence. *Geogr Anal* 33(3):195–214

- Rey SJ (2004) Spatial dependence in the evolution of regional income distributions. In: Getis A, Múr J, Zoeller H (eds) *Spatial econometrics and spatial statistics*. Palgrave, Hampshire, pp 193–213
- Rey SJ, Le Gallo J (2009) Spatial analysis of economic convergence. In: Mills TC, Patterson K (eds) *Handbook of econometrics volume II: applied econometrics*. Palgrave Macmillan, New York
- Rey SJ, Mack E, Koschinsky J (2012) Exploratory space–time analysis of burglary patterns. *J Quant Criminol* 28:509–531
- Rey SJ, Murray AT, Anselin L (2011) Visualizing regional income distribution dynamics. *Lett Spatial Resour Sci* 4(1):81–90
- Wikle C, Cressie N (2000) Space-time statistical modeling of environmental data. In Mowrer HT, Congalton RG (eds) *Quantifying spatial uncertainty in natural resources*. Ann Arbor Press, Chelsea, pp 213–235
- Williamson J (1965) Regional inequality and the process of national development. *Econ Dev Cult Change* 13(4):3–47
- Yang L, Xian G, Klaver J, Deal B (2003) Urban land-cover change detection through sub-pixel imperviousness mapping using remotely sensed data. *Photogramm Eng Rem Sens* 69(9):1003–1010

Eric M. Delmelle

Contents

70.1	Introduction	1386
70.1.1	Context	1386
70.1.2	One-Dimensional Sampling	1386
70.1.3	Two-Dimensional Sampling	1386
70.2	Geostatistical Sampling	1388
70.2.1	Designs for Variogram Estimation	1391
70.2.2	Optimal Designs to Minimize the Kriging Variance	1392
70.2.3	Sampling in a Multivariate Context	1393
70.3	Second-Phase Sampling	1393
70.4	Numerical Example	1395
70.5	Search Strategies	1397
70.6	Conclusion	1397
	References	1398

Abstract

Spatial sampling is the process of collecting observations in a two-dimensional framework. Careful attention is paid to (1) the quantity of the samples, dictated by the budget at hand, and (2) the location of the samples. A sampling scheme is generally designed to maximize the probability of capturing the spatial variation of the variable under study. Once initial samples have been collected and its variation documented, additional measurements can be taken at other locations. This approach is known as second-phase sampling, and various optimization criteria have recently been proposed to determine the optimal location of these

E.M. Delmelle

Department of Geography and Earth Sciences, University of North Carolina at Charlotte,
Charlotte, NC, USA

e-mail: Eric.Delmelle@uncc.edu

new observations. In this chapter, we review fundamentals of spatial sampling and second-phase designs. Their characteristics and merits under different situations are discussed, while a numerical example illustrates a modeling strategy to use covariate information in guiding the location of new samples. The chapter ends with a discussion on heuristic methods to accelerate the search procedure.

70.1 Introduction

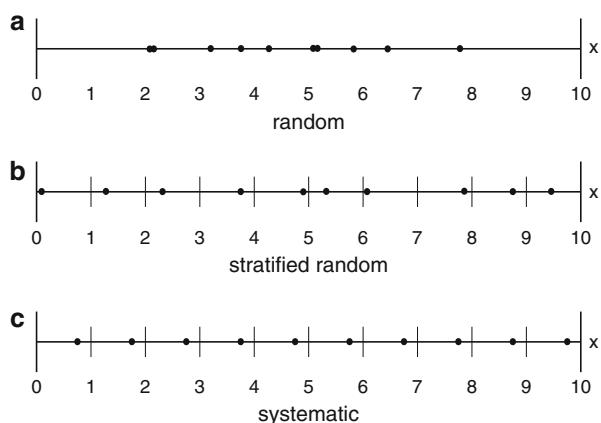
70.1.1 Context

According to Haining (2003), spatial or two-dimensional sampling has been applied to many disciplines such as mining, soil studies, telecommunications, ecology, geology, and geography, to cite a few (Akella et al. 2011). Scientists may be constrained by available budget and time to acquire a certain number of samples instead of trying to obtain information everywhere (Müller 1998; Thompson 2002; Delmelle 2009). It is generally desirable to find samples that are as representative as possible from the real data. Not only the cost of a complete census is prohibitive, it is time-consuming (Haining 1990) and it may result in redundant data when those are spatially autocorrelated (Griffith 2005). The autocorrelation function is defined as the similarity of the values of the variable of interest as a function of their separating distance (Gatrell 1979). This similarity decreases as the distance among sample points increases. Positive autocorrelation occurs when nearby observations are more alike than samples collected further away. Sparse sampling is less costly, but the variability of the variable of interest may go unnoticed. Consequently, not only the quantity of the samples is important but also their locations.

70.1.2 One-Dimensional Sampling

Pioneering research on sampling was devoted to one-dimensional problems (see, e.g., Cochran 1946; Madow 1946, 1953; Madow and Madow 1949). Cochran documented the efficiency associated with random sampling, systematic sampling, and stratified sampling. A *random sampling* scheme (Fig. 70.1a) allocates n sample points randomly within a population of interest. Each location is equally likely selected. In a *systematic random sampling* (Fig. 70.1b), the population is partitioned into a prespecified number of intervals. For each interval, a number of samples are collected, and the total of all samples is of size n . In a *systematic sampling* scheme (Fig. 70.1c), the population of interest is divided into n intervals of similar size. The first element is chosen within the first interval, starting at the origin, and the remaining $n - 1$ elements are aligned according to the same, fixed interval. A discussion of these configurations to the field of natural resources can be found in Stevens and Olsen (2004).

Fig. 70.1 One-dimensional sampling schemes for $n = 10$. The x -axis is partitioned in 10 intervals for cases (b) and (c). The random sampling locations have been generated using MATLAB rand function



70.1.3 Two-Dimensional Sampling

Necessary and common to both spatial and nonspatial sampling strategies are (i) the size of the sampling set, which is dictated by the budget (or time) at hand; (ii) the configuration of the sampling design; (iii) an estimator to characterize the population; and (iv) an estimation of the sampling variance to compute confidence intervals. Das (1950) has documented the variation of the sampling variance of two-dimensional designs. A *simple random sampling* design (Fig. 70.2a) randomly selects m sample points in a study region, generally denoted \mathfrak{D} , where each location has an equal opportunity to be sampled. In a *systematic sampling* design, (illustrations given in Fig. 70.2b–d), the study region is discretized into m intervals of equal size Δ . The first element is randomly or purposively chosen within the first interval, and so are other points in the remaining regions. If the first sample is chosen at random, the resulting scheme is called *systematic random sampling*. When the first sample point is not chosen at random, the resulting configuration is called *regular systematic sampling*. A *centric systematic sampling* occurs when the first point is chosen in the center of the first interval, resulting in a checkerboard configuration. The most common regular geometric configurations are the equilateral triangular grid, the rectangular (square) grid, and the hexagonal one (Cressie 1991). The benefits of a systematic approach reside in a good spreading of observations across \mathfrak{D} , guaranteeing a maximized sampling coverage, and preventing sampling clustering and redundancy. This design however presents two inconveniences:

- The distribution of separating distances in \mathfrak{D} is not represented well because many pairs of points are separated by the same distance,
- If the spatial process shows evidence of recurrence, periodicities, there is a risk that the variation of the variable will remain uncaptured, because the systematic design coincides in frequency with a regular pattern in the landscape (Overton and Stehman 1993).

A *systematic random method* addresses the second concern since it combines both systematic and random procedures (Dalton et al. 1975). One observation is

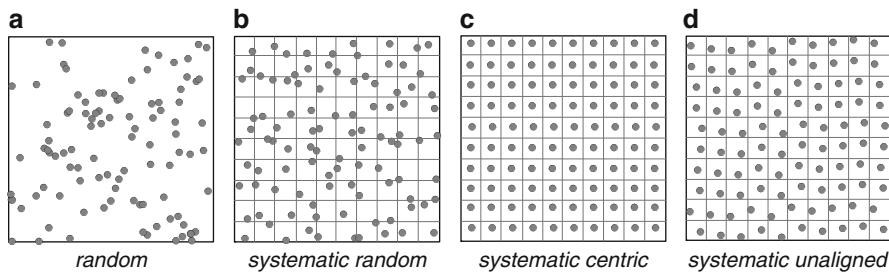


Fig. 70.2 Two-dimensional sampling schemes for $n = 100$. In figures (b), (c), and (d), both x - and y -axis have been divided into 10 intervals. Points were randomly generated using MATLAB rand function

randomly selected within each cell. However, sample density needs to be high enough in order to document the strength of the spatial relationship (e.g., variogram) among observations. From Fig. 70.2b, some patches of \mathfrak{D} remain undersampled, while others regions show evidence of clustered observations. A *systematic unaligned* scheme prevents this problem from occurring by imposing a stronger restriction on the random allocation of observations (King 1969).

In *stratified sampling* (Delmelle 2009), the population (or \mathfrak{D}) is partitioned into nonoverlapping strata. A set of samples is collected for each stratum, where the sum of the samples over all strata must equal m (strata may be of different size, for instance, a census tract). The knowledge of the underlying process is a determining factor in defining the shape and size of each stratum. Smaller strata are preferred in nonhomogeneous subregions.

Evaluation of Sampling Strategies. Following Quenouille's approach of a linear autocorrelation model, stratified random sampling is generally considered to yield a smaller variance than a systematic design. However, if the autocorrelation function is not linear (for instance, exponential), systematic sampling is the most efficient technique, followed by stratified random sampling and random sampling. Overton and Stehman (1993) presented some numerical results illustrating the magnitude of the differences of the three aforementioned designs under various population models. When sampling a phenomenon characterized by a regular pattern in the landscape, a systematic unaligned configuration is generally preferred (Delmelle 2009).

70.2 Geostatistical Sampling

An essential commonality of many natural phenomena is its spatial continuity in the geographical space. The field of geostatistics (Matheron 1963) provides a set of regression techniques to mathematically summarize the spatial variation of the phenomenon and use this information to predict the phenomenon under study at unsampled locations. Central to geostatistics is kriging, an interpolation technique that uses the semivariogram, a function which reflects the dissimilarity of pairs of points at different distance lags. The strength of this correlation determines the

weighting scheme used to create a prediction surface at unsampled locations, while minimizing the estimation error. As the distance separating two sample points increases, their similarity decreases and the influence on the weighting scheme diminishes. Beyond a specific distance called the range where autocorrelation is very small, the semivariogram flattens out (see, e.g., Ripley (1981) and Cressie (1991) for various summaries).

Mathematical Expression for Kriging. A variable of interest Y is collected at m supports within a study region \mathfrak{D} . Using data values of the primary variable, an empirical semivariogram $\hat{\gamma}(h)$ summarizes the variance of values separated by a particular distance lag (h):

$$\hat{\gamma}(h) = \frac{1}{2d(h)} \sum_{|\mathbf{s}_i - \mathbf{s}_j|=h} (y(\mathbf{s}_i) - y(\mathbf{s}_j))^2 \quad (70.1)$$

where $d(h)$ is the number of pairs of points for a given lag value, and $y(\mathbf{s}_i)$ the observation value at location \mathbf{s}_i . The semivariogram is characterized by a nugget effect a and a sill σ^2 where $\hat{\gamma}(h)$ levels out. The nugget effect reflects the spatial dependence at microscales, caused by measurement errors at distances smaller than sampling distances (Cressie 1991). Once the lag distance exceeds the range r , there is no spatial dependence between the sample sites anymore. The semivariogram function $\hat{\gamma}(h)$ becomes constant at a value called the sill σ^2 . A model $\gamma(h)$ is fitted to the experimental variogram, for instance, an exponential model:

$$\gamma(h) = \sigma^2 \left(1 - e^{-\frac{3h}{r}} \right) \quad (70.2)$$

In the presence of a nugget effect a , Eq. (70.2) becomes

$$\gamma(h) = a + (\sigma^2 - a) \left(1 - e^{-\frac{3h}{r}} \right) \quad (70.3)$$

Equation (70.4) denotes the corresponding covariogram $C(h)$ that summarizes the covariance between any two points:

$$C(h) = C(0) - \gamma(h) = \sigma^2 - \gamma(h) \quad (70.4)$$

The interpolated, kriged value at a location \mathbf{s} in space is given as a weighted mean of surrounding values, where each value is weighted according to the variogram model:

$$\hat{y}(\mathbf{s}) = \sum_{i=1}^W w_i(\mathbf{s}) y(\mathbf{s}_i) \quad (70.5)$$

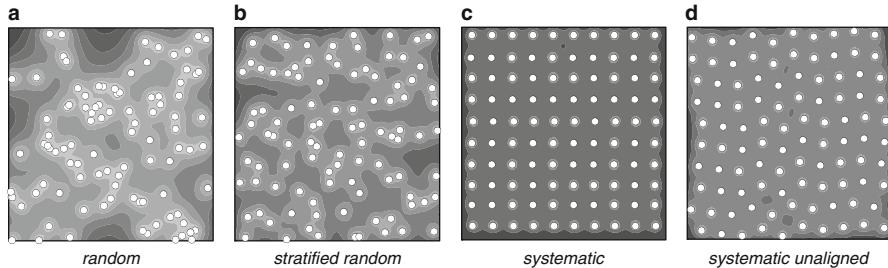


Fig. 70.3 Kriging variance associated with the two-dimensional sampling schemes of Fig. 70.3

where W is the set of neighboring points that are used to estimate the interpolated value at location \mathbf{s} , and $w_i(\mathbf{s})$ is the weight associated with each surrounding point, which is a function of the semivariogram function. The weight of each sample can be determined by an exponential function (Eq. (70.2)). For computational purposes, kriging is performed on a set of grid nodes \mathbf{s}_g ($g = 1, 2, \dots, G$). Kriging yields an associated variance that measures the prediction uncertainty. The kriging variance at a location \mathbf{s} is given by

$$\sigma_k^2(\mathbf{s}) = \sigma^2 - \mathbf{c}^T(\mathbf{s})\mathbf{C}^{-1}\mathbf{c}(\mathbf{s}) \quad (70.6)$$

where \mathbf{c}^T is the transpose of the covariance matrix \mathbf{C} based on the covariogram function and \mathbf{C}^{-1} its inverse. The overall kriging variance (σ_k^2) is obtained by integrating Eq. (70.6) over the region \mathfrak{D} . Computationally, it is easier to perform a spatial of \mathfrak{D} and sum the kriging variance over all grid points \mathbf{s}_g :

$$\int_{\mathfrak{D}} \sigma_k^2(\mathbf{s}_g) \approx \frac{1}{G} \sum_{g \in G} \sigma_k^2(\mathbf{s}_g) \quad (70.7)$$

The kriging variance can be calculated with an estimated variogram and the known location of existing sampling points. The kriging variance solely depends on the spatial dependence and configuration of the observations (Cressie 1991). Figure 70.3 summarizes the variation in the kriging variance estimate for the four designs of Fig. 70.2.

Van Groenigen et al. (1998, 1999) suggest that initial sampling schemes should be optimized for a reliable estimation of the variogram function, which can either be used for the prediction of the variable under study or to help designing additional sampling phase(s). For the former, two strategies have been suggested in the literature:

- (a) A geometric coverage of sample points over the study region is generally desirable to guarantee enough pairs of points at different distances.
- (b) Points need to be distributed in the multivariate field to capture as much variation as possible.

Moreover, optimal sampling strategies exist to reduce the kriging variance associated with the interpolation process. The next paragraphs illustrate three common objectives in spatial sampling: variogram estimation, minimization of the kriging variance, and sampling in a multivariate field.

70.2.1 Designs for Variogram Estimation

Traditional ways to evaluate the goodness of a sampling scheme do not incorporate the spatial structure of the variable. The increasing use of geostatistics as a least-squares interpolation technique, however, has fostered research on optimizing sampling configurations to maximize the amount of information obtained during a first sampling phase. Matérn (1960) and Yfantis et al. (1987) have suggested that the use of an equilateral triangular sampling grid (Fig. 70.4) can yield to a very reliable estimation of the variogram and predict the mean over a study region, assuming radially symmetric, decreasing covariances.

Systematic designs (Fig. 70.2c, d) offer the advantage of good coverage of observations, capturing the main features of the variogram (Van Groenigen et al. 1999). It may be necessary to strategically design a scheme where a subset of the observations are evenly spread across the study area and the remaining points clustered together to capture the autocorrelation function at very small distances (Delmelle 2009).

The reliability of the variogram function depends on the number of pairs of points within each distance band. Russo (1984) and Warrick and Myers (1987) have proposed some strategies to reproduce an a priori defined ideal distribution of pairs of points, based on a given variogram function. The procedure allows to account for the variation in distance and direction (anisotropy¹). Corsten and Stein (1994) use a nested sampling design for a better estimation of the nugget effect. A nested sampling design consists of taking observations according to a hierarchical scheme, with decreasing distances between observations. This type of sampling scheme distributes a high number of observations in some parts of the area and a low observation density in other regions. This in turn generates only a few distances for which variogram values are available. Taking into account a prior information of the spatial structure of the variable and assuming a stationary variable, Van Groenigen and Stein (1998) have combined two different objectives to allocate samples during an initial phase. The first objective called the Warrick/Myers criterion ensures optimal estimation of the covariogram and aims at redistributing pairs of points over the distance and direction lags according to a prespecified distribution. The second criterion, called *minimization of the mean of the shortest distances* (MMSD), requires all sampling points spread evenly to ensure that

¹Anisotropy is a property of a natural process, where the autocorrelation among points changes with the distance and direction between two locations. We talk about an isotropic process however when there is no effect of direction in the spatial autocorrelation of the primary variable.

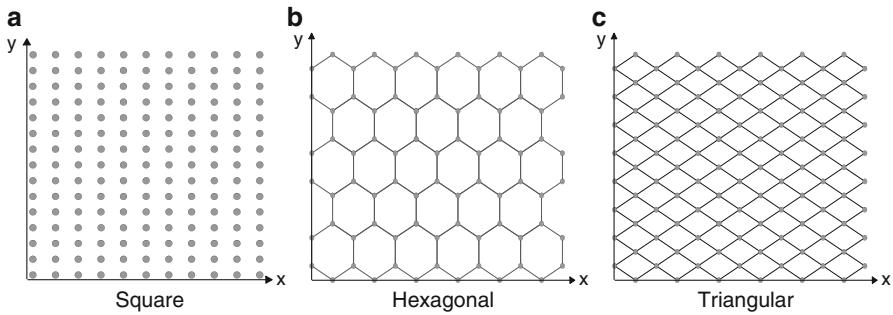


Fig. 70.4 Three common geometric sampling schemes

unsampled locations are never far from a sampling point. The second criterion suggested by the authors is of deterministic nature, resulting an even spreading pairs of points across the study area, which is similar in nature to a systematic pattern.

70.2.2 Optimal Designs to Minimize the Kriging Variance

The kriging procedure generates a minimum-error estimate of the variable of interest. This uncertainty is minimal – or zero when there is no nugget effect – at existing sampling points and increases with the distance to the nearest samples. One approach suggested in the literature is to design a sampling configuration to minimize this uncertainty. Since continuous sampling is not feasible, seeking the best sampling procedure must be carried out on a discretized grid. Using an a priori variogram model (Eq. (70.4)), it is possible to design an initial sampling scheme S to minimize the overall kriging variance (Eq. (70.8)) or the maximum kriging variance (Eq. (70.9)).

$$\underbrace{\text{MINIMIZE}}_{\{s_1, \dots, s_m\}} J(S) = \frac{1}{G} \sum_{g \in G} \sigma_k^2(s_g; S) \quad (70.8)$$

$$\underbrace{\text{MINIMIZE}}_{\{s_1, \dots, s_m\}} J(S) = \frac{1}{G} \underbrace{\sup}_{g \in G} \{ \sigma_k^2(s_g; S) \} \quad (70.9)$$

Burgess et al. (1981) estimate kriging variances for different scenarios of sampling densities, nugget effects, and size of study regions. The strategy attempts to identify the minimum number of samples necessary to reach a certain level of kriging variance. General findings are the increase of the prediction error as the nugget effect increases or when the sampling density is reduced. An equilateral

triangular configuration of sampling points (Fig. 70.4c) is best under isotropic conditions, but a square grid at the same density is nearly as good, and is preferred for data collection convenience. An equilateral triangle design will keep the variance to a minimum, because it reduces the farthest distance from initial sample points to nonsample points. A square grid performs well, especially in case of isotropy (McBratney and Webster 1981; McBratney et al. 1981). When directional discontinuities are present, a square grid pattern is preferred (Olea 1984; Yfantis et al. 1987).

70.2.3 Sampling in a Multivariate Context

McBratney and Webster (1983) have discussed the importance of spatial sampling to multivariate fields. Sample data can be very difficult to collect, and very expensive, especially when monitoring air or soil pollution, for instance (Haining 1990). Secondary data can be a valuable asset if they are available continuously over a study area and combined within the primary variable (Hengl et al. 2003). Secondary spatial data sources can include maps, digital elevation models, and national, socioeconomic, and demographic census data. Cross-variograms express the spatial relationships among those variables. In turn, this information is capitalized to calibrate the parameters of the kriging equations. When the variogram of the primary variable and the cross-variograms are known a priori, an improved sampling configuration can be obtained. A rule of thumb consists of locating the observations of the main variable where covariates exhibit substantial spatial variation (Delmelle and Goovaerts 2009). Secondary variables should be used to reduce the sampling effort in areas where their local contribution in predicting the primary variable is maximum (Delmelle 2009). If a set of covariates predicts accurately the data value where no initial sample has been collected yet, there is little incentive to perform sampling at that location. On the other hand, when covariates perform poorly in estimating the primary variable, additional samples are necessary.

70.3 Second-Phase Sampling

Second-phase spatial sampling is defined as the addition of new observations to improve the overall prediction of the variable of interest. A set M of m initial measurements has been collected, and a variogram that summarizes the spatial structure of the variable of interest will help to determine the location and size for an additional set and location of their samples. It is generally agreed in the literature that the objective function aims to collect new samples to reduce the prediction error (kriging variance) by as much as possible.

Mathematical Expression for Minimizing of the Kriging Variance in a Second Phase. We add a set of n new sample points to the initial sample set of size m .

Using the variogram function from the first sample set, the change in kriging variance $\Delta\sigma_k^2$ is over all grid points \mathbf{s}_g :

$$\Delta\sigma_k^2 = \frac{1}{G} \left[\sum_{g \in G} \sigma_k^{\text{old}}(\mathbf{s}_g) - \sum_{g \in G} \sigma_k^{\text{new}}(\mathbf{s}_g) \right] \quad (70.10)$$

where σ_k^{old} is the mean kriging variance calculated with the set of $[m]$ initial sample points and σ_k^{new} is the mean kriging variance with the $[m+n]$ additional set of points. From Eq. (70.10)

$$\sigma_k^{\text{old}}(\mathbf{s}_g) = \sigma^2 - \underbrace{\mathbf{c}(\mathbf{s}_g)}_{[1,m]} \times \underbrace{\mathbf{C}^{-1}}_{[m]} \times \underbrace{\mathbf{c}^T(\mathbf{s}_g)}_{[m,1]} \quad (70.11)$$

$$\sigma_k^{\text{new}}(\mathbf{s}_g) = \sigma^2 - \underbrace{\mathbf{c}(\mathbf{s}_g)}_{[1,m+n]} \times \underbrace{\mathbf{C}^{-1}}_{[m+n]} \times \underbrace{\mathbf{c}^T(\mathbf{s}_g)}_{[m+n,1]} \quad (70.12)$$

The objective function helps to locate the set of additional n points that will maximize this change in kriging variance (Christakos and Olea 1992; Van Groenigen et al. 1999; Rogerson et al. 2004). The n additional points are to be chosen from a set of size $(N-m)$, that is, all possible sample sites in \mathfrak{D} except the m ones selected during the first sampling phase. In that case, there are $\binom{N-m}{n}$ possible combinations and it is almost impossible to find the optimal using. The objective function is formulated as follows:

$$\underbrace{\text{MAXIMIZE}_{\{\mathbf{s}_{m+1}, \dots, \mathbf{s}_{m+n}\}}} J(S) = \frac{1}{G} \sum_{g \in G} \Delta\sigma_k^2(s_g; S) \quad (70.13)$$

Incorporating Secondary Information in a Second Sampling Phase. New samples can be collected in areas where secondary variables do not provide good estimates of the primary variable. Consider the situation where the primary data is supplemented by k additional secondary variables X_i ($\forall i = 1, \dots, k$) available at G grid nodes \mathbf{s}_g ($g = 1, 2, \dots, G$). Local regression techniques such as geographically weighted regression (Brunsdon et al. 1996) provide locally linear regression estimates at every point i , using distance weighted samples. Our goal is to sample in those areas characterized by low local r^2 , since it is in those areas that covariates are not performing well in predicting the outcome of the primary variable. A local r^2 can be conceived as how well covariates predict the main variable locally, for instance, from a GWR model.

Formulating the Second-Phase Sampling Problem. This approach is proposed by Cressie and has been applied by Rogerson et al. (2004) and Delmelle and

Goovaerts (2009) to weight the kriging variance, where the importance of a location to be sampled is represented by a weight $w(\mathbf{s})$, which is location specific.

$$\underbrace{\text{MAXIMIZE}_{\{s_{m+1}, \dots, s_{m+n}\}}}_{J(S)} J(S) = \frac{1}{G} \sum_{g \in G} w(s_g) \Delta \sigma_k^2(s_g; S) \quad (70.14)$$

The weight should reflect the importance provided locally by covariates, but could also account for the rapid change in spatial structure of the primary variable at s_g (Delmelle and Goovaerts 2009).

70.4 Numerical Example

A numerical example is provided to gain insight into the structure of the sampling problem. The goal is to maximize the change in the weighted kriging variance. As a hypothetical example, we simulated a synthetic snowfall data in a 10×10 km bounding box.

Minimizing the Kriging Variance. Figure 70.5 displays the initial set of 50 measurements and the associated interpolated map, based on an exponential semivariogram with a range of 3,000 m, a nugget effect $C(0) = 0$, and sill $a = 0.025$. The amount of snowfall is simulated to be minimal in the upper northwestern corner and increases steadily southeastwards. Figure 70.6 on the left is an interpolated contour map of the prediction error. The variance increases away from existing data points, to reach maximum values in the corner of the study area. The right figure displays the discretized study area, generating a set $P = 51 \times 51$ potential points. If the goal is to maximize the change in kriging variance only (Eq. (70.13)), the location of the new points would be far away from existing ones.

Weights to Reflect Sampling Priorities. An example of sampling weights is given in Fig. 70.3 on the left and is multiplied by the kriging variance on the right. As a result of this multiplication, some locations exhibiting high weight for second-phase sampling (where we observe a stronger variation in the spatial structure of the primary variable) may not be recommended for further sampling since they are located in the close vicinity of an existing initial sampling point. For instance, on the left figure, the location [6, 705, 000; 4, 718, 500] is characterized by a high sampling weight. However, the region has already been sampled and consequently the likelihood for second-phase sampling decreases. Consider that we have the intention to add one sample point that would optimize Eq. (70.14). To find the point that would maximize the change in kriging variance, summed over all grid points, an iterative procedure is necessary which evaluates the score of each candidate sample points to the objective function. Such an enumeration can be very time-consuming, and for computational purposes, it is less demanding to select a point where the weighted kriging variance value is maximum. From Fig. 70.7, the location of the point exhibiting the maximum weighted kriging variance was

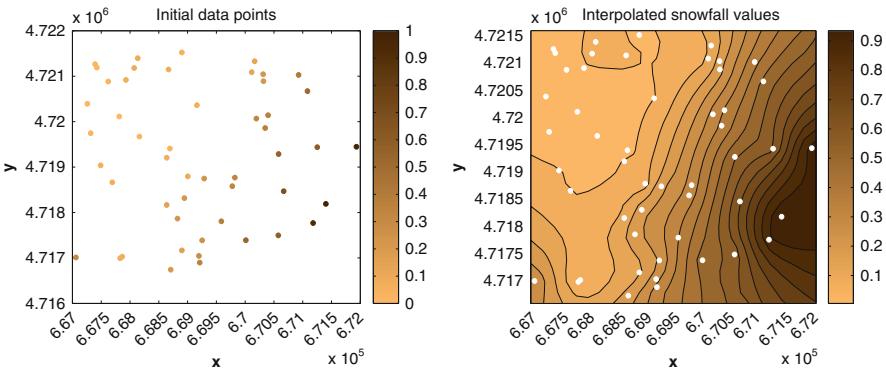


Fig. 70.5 Dark colors denote regions characterized by heavier amounts of snow. Units are in feet

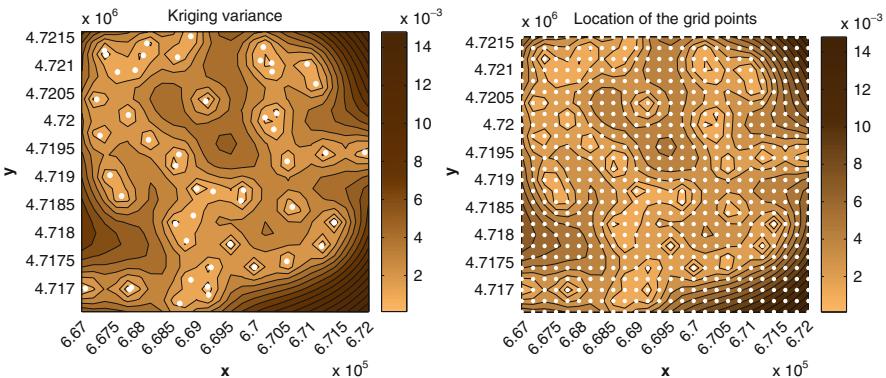


Fig. 70.6 The prediction error on the left and the location of grid nodes s_g

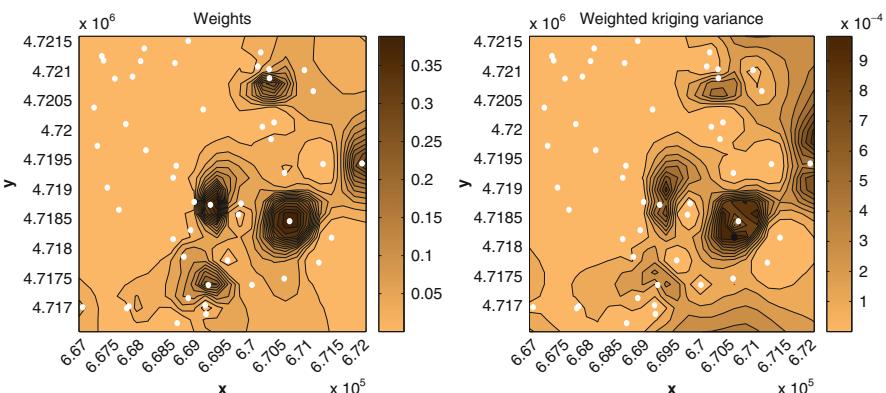


Fig. 70.7 Combined objectives 2 and 3 (on the left), multiplied with the kriging variance (on the right)

[6, 70, 600; 4, 718, 200]. Once an optimal (or near optimal) point has been added, it is possible to recompute the objective function. It is also desirable to adapt the constraints as the iteration continues.

70.5 Search Strategies

The set of candidate sampling locations may be large, and it is desirable to rely on heuristic techniques to return an acceptable solution in a limited time frame. A heuristic guides the search towards a sample set S that is optimal (or near optimal) to a predefined objective function, for instance, the set S^* is optimal to the objective function J defined in Eq. (70.14). The efficiency of a heuristic depends on its capacity to give as often as possible a solution close to S^* . In second-phase sampling, a heuristic that would select m points at random would not return a very good value for J . Examples of search techniques include the *greedy* algorithm, *simulated annealing*, *tabu search*, and *genetic algorithms*, among others (Christakos and Olea 1992; Delmelle and Goovaerts 2009). The *greedy* algorithm builds a solution sequentially only accepting improving moves, while the other methods improve the value of the objective function by iterations starting from an initial solution s_0 , but also accepting non-improving moves. Specifically, the three first methods usually remain stuck in a local optimum while the last three – also called *metaheuristics* – may find the optimal solution s^* . Greedy leads to a unique solution S_0^+ and does not explore the entire set of candidate samples. *Simulated annealing* authorizes to occasionally decrease the objective function (in our case to be maximized) in order to continue exploring for better solutions. Note that the simulated annealing algorithm does not always converge. Both *tabu search* and *genetic algorithm* techniques have not been applied to sampling optimization. They both lead to an optimal solution, yet the tabu search algorithm tends to cycle.

70.6 Conclusion

Accurate and effective spatial sampling strategies are very important when researchers are limited by their available budget (or time). A careful design is crucial to identify the main features of the phenomenon under study and avoid that its spatial characteristics remain unnoticed. For instance, incorporating some randomness in a systematic sampling design may be useful to document patterns with periodicities. Once initial samplings have been collected, a variogram can be built, which ultimately helps designing a second-phase sampling survey (away from existing samples and where the variation is maximum). When the set of candidate locations is large and the objective nonlinear, heuristic methods may be necessary to find a set optimal to some sampling criteria. The methods illustrated in this chapter may easily be extended to areal data (for instance, census tracts or socio-economic strata). Some areas may be deemed more important for sampling, and the proposed objectives are flexible to reflect sampling priorities.

References

- Akella MR, Delmette EM, Batta R, Rogerson P, Blatt A (2011) Adaptive cell tower location using geostatistics. *Geogr Anal* 42(3):227–244
- Brunsdon C, Fotheringham AS, Charlton ME (1996) Geographically weighted regression: a method for exploring spatial nonstationarity. *Geogr Anal* 28(4):281–298
- Burgess TM, Webster R, McBratney AB (1981) Optimal interpolation and isarithmic mapping of soil properties: IV. Sampling strategy. *J Soil Sci* 32(4):643–659
- Christakos G, Olea RA (1992) Sampling design for spatially distributed hydrogeologic and environmental processes. *Adv Water Res* 15(4):219–237
- Cochran WG (1946) Relative accuracy of systematic and stratified random samples for a certain class of populations. *Ann Math Stat* 17(2):164–177
- Corsten LCA, Stein A (1994) Nested sampling for estimating spatial semivariograms compared to other designs. *Appl Stoch Models Data Anal* 10(2):103–122
- Cressie N (1991) Statistics for spatial data. Wiley, New York
- Dalton R, Garlick J, Minshull R, Robinson A (1975) Sampling techniques in geography. Goerges Philip and Son, London
- Das AC (1950) Two-dimensional systematic sampling and the associated stratified and random sampling. *Sankhyā* 10(1–2):95–108
- Delmelle E (2009) Spatial sampling. In: Rogerson P, Fotheringham S (eds) *The SAGE handbook of spatial analysis*. SAGE London
- Delmelle E, Goovaerts P (2009) Second-phase spatial sampling designs for non-stationary spatial variables. *Geoderma* 153(1–2):205–216
- Gatrell AC (1979) Autocorrelation in spaces. *Environ Plan A* 11(5):507–516
- Griffith DA (2005) Effective geographic sample size in the presence of spatial autocorrelation. *Ann Assoc Am Geogr* 95(4):740–760
- Haining RP (1990) Spatial data analysis in the social and environmental sciences. Cambridge University Press, Cambridge, UK
- Hengl T, Rossiter DG, Stein A (2003) Soil sampling strategies for spatial prediction by correlation with auxiliary maps. *Aust J Soil Res* 41(8):1403–1422
- King LJ (1969) Statistical analysis in geography. Prentice-Hall, Englewood Cliffs, pp 217–222
- Madow LH (1946) Systematic sampling and its relation to other sampling designs. *J Am Stat Assoc* 41(234):204–217
- Madow WG (1953) On the theory of systematic sampling. III. Comparison of centered and random start systematic sampling. *Ann Math Stat* 24(1):101–106
- Madow WG, Madow LH (1949) On the theory of systematic sampling. I. *Ann Math Stat* 15(1):1–24
- Matérn B (1960) Spatial variation. Springer, Berlin/Heidelberg/New York, p 151
- Matheron G (1963) Principles of geostatistics. *Econ Geol* 58(8):1246–1266
- McBratney AB, Webster R (1981) The design of optimal sampling schemes for local estimation and mapping of regionalized variables: II. Program and examples. *Comput Geosci* 7(4):331–334
- McBratney AB, Webster R (1983) Optimal interpolation and isarithmic mapping of soil properties: V. Co-regionalization and multiple sampling strategy. *J Soil Sci* 34(1):137–162
- McBratney AB, Webster R, Burgess TM (1981) The design of optimal sampling schemes for local estimation and mapping of regionalized variables: I. Theory and method. *Comput Geosci* 7(4):335–365
- Müller W (1998) Collecting spatial data: optimal design of experiments for random fields. Physica, Heidelberg
- Olea RA (1984) Sampling design optimization for spatial functions. *Math Geol* 16(4):369–392
- Overton WS, Stehman SV (1993) Properties of designs for sampling continuous spatial resources from a triangular grid. *Commun Stat Theory Methods* 22(9):2641–2660
- Ripley BD (1981) Spatial statistics. Wiley, New York, p 252

- Rogerson PA, Delmelle EM, Batta R, Akella MR, Blatt A, Wilson G (2004) Optimal sampling design for variables with varying spatial importance. *Geogr Anal* 36(2):177–194
- Russo D (1984) Design of an optimal sampling network for estimating the variogram. *Soil Sci Soc Am J* 48(4):708–716
- Stevens D, Olsen A (2004) Spatially balanced sampling of natural resources. *J Am Stat Assoc* 99(465):262–278
- Thompson SK (2002) Sampling, 2nd edn. Wiley, New York, p 367
- Van Groenigen JW, Stein A (1998) Constrained optimization of spatial sampling using continuous simulated annealing. *J Environ Qual* 27(5):1078–1086
- Van Groenigen JW, Siderius W, Stein A (1999) Constrained optimisation of soil sampling for minimisation of the kriging variance. *Geoderma* 87(3–4):239–259
- Warrick AW, Myers DE (1987) Optimization of sampling locations for variogram calculations. *Water Resour Res* 23(3):496–500
- Yfantis EA, Flatman GT, Behar JV (1987) Efficiency of kriging estimation for square, triangular and hexagonal grids. *Math Geol* 19(3):183–205

Spatial Models Using Laplace Approximation Methods

71

Virgilio Gómez-Rubio, Roger S. Bivand, and Håvard Rue

Contents

71.1	Introduction	1402
71.2	Integrated Nested Laplace Approximation	1403
71.2.1	Gaussian Markov Random Fields	1405
71.2.2	Priors	1406
71.2.3	Model Criticism and Selection	1407
71.2.4	Implementation	1408
71.2.5	Other Features	1408
71.3	Spatial Models	1408
71.3.1	Geoadditive Mixed-Effects Models	1408
71.3.2	Disease Mapping	1409
71.3.3	Geostatistical Models	1411
71.3.4	Point Process Models	1412
71.4	Examples	1413
71.4.1	Geostatistics	1413
71.4.2	Lattice Data	1413
71.4.3	Point Patterns	1413
71.5	Conclusions	1415
	References	1416

V. Gómez-Rubio (✉)

Department of Mathematics, School of Industrial Engineering-Albacete, University of Castilla-La Mancha, Albacete, Spain

e-mail: virgilio.gomez@uclm.es

R.S. Bivand

Department of Economics, NHH Norwegian School of Economics, Bergen, Norway

e-mail: Roger.Bivand@nhh.no

H. Rue

Department of Mathematical Sciences, Norwegian University of Science and Technology, Trondheim, Norway

e-mail: Havard.Rue@math.ntnu.no

Abstract

Bayesian inference has been at the center of the development of spatial statistics in recent years. In particular, Bayesian hierarchical models including several fixed and random effects have become very popular in many different fields. Given that inference on these models is seldom available in closed form, model fitting is usually based on simulation methods such as Markov chain Monte Carlo.

However, these methods are often very computationally expensive and a number of approximations have been developed. The integrated nested Laplace approximation (INLA) provides a general approach to computing the posterior marginals of the parameters in the model. INLA focuses on latent Gaussian models, but this is a class of methods wide enough to tackle a large number of problems in spatial statistics.

In this chapter, we describe the main advantages of the integrated nested Laplace approximation. Applications to many different problems in spatial statistics will be discussed as well.

71.1 Introduction

Spatial models provide a suitable way of analyzing data when observations are thought to be correlated because of their locations in space. Bayesian inference has proven useful when dealing with spatial models and modeling local dependence. In Bayesian analysis (see, e.g., Gelman et al. 2003), inference about the vector of model parameters \mathbf{x} is based on computing their joint posterior distribution given the vector of observed data \mathbf{y} . This is done by means of Bayes' rule:

$$\pi(\mathbf{x}|\mathbf{y}) \propto \pi(\mathbf{y}|\mathbf{x})\pi(\mathbf{x})$$

Here $\pi(\mathbf{y}|\mathbf{x})$ represents the likelihood of the model given its parameters and $\pi(\mathbf{x})$ is the prior distribution of the parameters of the model. Hence, the posterior distribution depends on the mechanism which generates the data (i.e., the likelihood) and the previous information about the model parameters (i.e., the prior distribution). Note that $\pi(\mathbf{x})$ is often supposed to depend on some hyperparameters which in turn have their own prior distributions.

$\pi(\mathbf{x}|\mathbf{y})$ is a multivariate distribution of the ensemble of model parameters which is often hard to obtain. In many applications it is sufficient with obtaining a separate posterior distribution for some of the parameters in the model because no joint inference is needed (e.g., the estimates of the relative risk in different areas). These distributions are called posterior marginals and can be denoted $\pi(x_i|\mathbf{y})$.

As these are univariate distributions, they are often easier to compute or approximate than the joint posterior distribution.

Given that in most cases there is no closed form for the posterior distributions of most parameters in the model, Markov chain Monte Carlo

(MCMC, see Gelman et al. 2003) techniques have been employed to estimate the joint posterior. Furthermore, a number of sound techniques for model criticism, comparison, and selection make Bayesian inference appealing.

For models with complex spatial dependence or large datasets, MCMC may not be a convenient solution due to computational time. For this reason, Rue et al. (2009) propose the use of approximate inference based on what they have called the integrated nested Laplace approximation (INLA). This approximation will focus on the posterior marginals which are easier to compute than obtaining an approximation to the joint posterior distribution. Also, INLA will only consider approximations for hierarchical models whose latent effects can be expressed as a Gaussian Markov random field (GMRF).

Successful applications of INLA include disease mapping (Schroedle et al. 2011), geostatistics (Eidsvik et al. 2009), point patterns (Illian et al. 2012), and others (Martino and Rue 2010).

71.2 Integrated Nested Laplace Approximation

The integrated nested Laplace approximation (INLA) focuses on providing a good approximation to the posterior marginal distributions of the parameters in the model. In particular, this approximation has been developed for latent Gaussian models. These cover a general class of models which appear in many areas of interest. Spatial statistics is one of them, as spatial correlation can be introduced by means of correlated random effects.

First of all, let us assume that we have n observed variables $y_i, i = 1, \dots, n$ with a distribution (usually from the exponential family) with a mean μ_i which is related to a linear predictor η_i through a convenient link function. In turn, η_i is modeled additively on different effects:

$$\eta_i = \alpha + \sum_{j=1}^{n_f} f^{(j)}(u_{ji}) + \sum_{k=1}^{n_\beta} \beta_k z_{ki} + \varepsilon_i$$

Here, $f^{(j)}$ represents some nonlinear function or random effects (of which there are n_f) on a set of covariates \mathbf{u} , β_k are coefficients for linear effects on a vector of covariates \mathbf{z} , and ε_i are unstructured terms. The latent effects $\mathbf{x} = \{\{\eta_i\}, \alpha, \{\beta_k\}, \dots\}$ are assumed to be Gaussian with zero mean and precision matrix $\mathbf{Q}(\theta_1)$, where θ_1 is a vector of hyperparameters. Hence, the observations will have a likelihood which will depend on the latent effects \mathbf{x} and a set of parameters θ_2 . Furthermore, the observations y_i are supposed to be independent given \mathbf{x} and θ_2 .

In the particular case of spatial statistics, the terms $f^{(j)}(u_{ji})$ can be taken as $f_i^{(j)}$ (or u_i abusing of notation) to represent a random effect at a spatial location i . Hence, covariate u_{ji} acts as the spatial index i of area i for the set of random effects j . For example, taking $n_f = 2$ we can define $u_i = f^1(u_{1i})$ and $v_i = f^2(u_{2i})$,

where $\mathbf{u} = \{u_1, \dots, u_n\}$ is a vector of independent random effects and $\mathbf{v} = \{v_1, \dots, v_n\}$ is a vector of spatially correlated random effects.

Rue et al. (2009) focus on the posterior distribution of \mathbf{x} and the vector of hyperparameters $\theta = (\theta_1, \theta_2)$:

$$\begin{aligned}\pi(\mathbf{x}, \theta | \mathbf{y}) &\propto \pi(\theta) \pi(\mathbf{x} | \theta) \prod_{i \in \mathcal{I}} \pi(y_i | x_i, \theta) \propto \\ \pi(\theta) |\mathbf{Q}(\theta)|^{n/2} \exp \left\{ -\frac{1}{2} \mathbf{x}^T \mathbf{Q}(\theta) \mathbf{x} + \sum_{i \in \mathcal{I}} \log(\pi(y_i | x_i, \theta)) \right\}\end{aligned}$$

Here \mathcal{I} is the subset of indices (from 1 to length of \mathbf{x} , the number of latent effects) that are observed with observations \mathbf{y} and their respective linear predictors $\{\eta_i\}$. Note that η_i is the only observed latent effect (through y_i) and that all the other latent effects are not observed directly and need to be estimated. In addition, the latent effects may be subject to some linear constraints of the form $\mathbf{Ax} = \mathbf{e}$. Finally, the latent field is supposed to have conditional independence properties, so that \mathbf{x} becomes a Gaussian Markov random field (GMRF). As we will show later, these Markov properties play an important role when modeling spatial data.

The likelihood of the data $\pi(\mathbf{y} | \mathbf{x}, \theta)$ is not constrained to be Gaussian. At the moment, INLA can deal with several likelihoods from the exponential family as well as with mixtures, such as zero-inflated distributions. Furthermore, INLA is flexible enough to allow different observations to have different likelihoods. Hence, INLA can deal with a myriad of models.

Instead of aiming at the full posterior distribution of the model parameters \mathbf{x} and θ , Rue et al. (2009) focus on obtaining an approximation to the posterior marginal distributions $\pi(x_i | \mathbf{y})$ and $\pi(\theta_j | \mathbf{y})$. These marginals can be written down as

$$\pi(x_i | \mathbf{y}) \propto \int \pi(x_i | \theta, \mathbf{y}) \pi(\theta | \mathbf{y}) d\theta$$

and

$$\pi(\theta_j | \mathbf{y}) \propto \int \pi(\theta | \mathbf{y}) d\theta_{-j}$$

Here θ_{-j} denotes θ minus component θ_j .

The approximations will be for the conditional distributions in the right-hand sides of the previous expressions. Note that an approximation to $\pi(\theta | \mathbf{y})$ is also required and that numerical integrations will be feasible only if the dimension of θ is small (as it often happens in practice).

A first approximation to $\pi(\theta | \mathbf{y})$ using Gaussian distributions can be constructed as follows:

$$\tilde{\pi}(\theta | \mathbf{y}) \propto \frac{\pi(\mathbf{x}, \theta, \mathbf{y})}{\tilde{\pi}_G(\mathbf{x} | \theta, \mathbf{y})} \Big|_{x=x^*(\theta)}$$

$\tilde{\pi}_G(\mathbf{x}|\theta, \mathbf{y})$ is the Gaussian approximation to the full conditional of \mathbf{x} and $x^*(\theta)$ is the mode of the full conditional for a given value of θ .

Hence, the marginals of interest can be computed using numerical integration over a multidimensional grid of values of θ . For example,

$$\tilde{\pi}(x_i|\mathbf{y}) = \sum_k \tilde{\pi}(x_i|\theta_k, \mathbf{y}) \times \pi(\theta_k|\mathbf{y}) \times \Delta_k$$

where Δ_k represents the weights for each vector of values θ_k in the grid.

Rue and Martino (2007) and Rue et al. (2009) stress the importance of having a good approximation to $\pi(x_i|\theta, \mathbf{y})$. A Gaussian approximation $\tilde{\pi}_G(x_i|\theta, \mathbf{y})$ is based on using a normal distribution with mean $\mu_i(\theta)$ and marginal variance $\sigma_i^2(\theta)$. The approximation provided by INLA (and in particular the Gaussian approximation for $\pi(\mathbf{x}|\theta, \mathbf{y})$) is exact for Gaussian data and the approximation is only due to integration (with respect to θ) error. This may be a good starting point, but it may not suffice because of possible inaccuracy if it is not centered at the correct point and because of its lack of skewness.

For this reason, they also propose other alternatives such as the Laplace approximation and the *integrated nested Laplace approximation* (INLA). Firstly, an improved approximation may be obtained by using a Laplace approximation:

$$\tilde{\pi}_{LA}(x_i|\theta, \mathbf{y}) \propto \frac{\pi(\mathbf{x}, \theta, \mathbf{y})}{\tilde{\pi}_{GG}(\mathbf{x}_{-i}|x_i, \theta, \mathbf{y})} |_{\mathbf{x}_{-i}=\mathbf{x}_{-i}^*(x_i, \theta)}$$

Here $\tilde{\pi}_{GG}(\mathbf{x}_{-i}|x_i, \theta, \mathbf{y})$ is a Gaussian approximation to $\mathbf{x}_{-i}|x_i, \theta, \mathbf{y}$ which is centered around the mode $\mathbf{x}_{-i}^*(x_i, \theta)$. As this approximation must be computed for every x_i , some numerical techniques are required to speed up computation.

Finally, Rue et al. (2009) derive a simplified Laplace approximation to improve the approximation given by $\tilde{\pi}_{LA}(x_i|\theta, \mathbf{y})$ by means of a series expansion of the Laplace approximation around $x_i = \mu_i(\theta)$. This provides a better approximation and it corrects for location and skewness. As $\tilde{\pi}_{LA}(x_i|\theta, \mathbf{y})$ is very expensive to compute, the simplified Laplace approximation seems the best trade-off between speed and accuracy.

It should be noted that while these approximations will center on the posterior marginal of a single latent effect x_i or hyperparameter θ_i , the methodology behind them could be applied to obtain an approximation of the joint posterior of any subset S of latent effects \mathbf{x}_S (see Sect. 6.1, Rue et al. 2009). However, in that case, the approximations become more complex and the numerical integration needed is more demanding.

71.2.1 Gaussian Markov Random Fields

Approximate inference using INLA is based on the assumption that the latent field \mathbf{x} is Gaussian and fulfills some conditional independence properties. In particular,

any two latent effects x_i and x_j in \mathbf{x} should be independent given the remaining latent effects \mathbf{x}_{-ij} . Furthermore, the number of hyperparameters appearing in the distribution of \mathbf{x} is assumed to be small.

Rue and Held (2005) provide a description of methods for efficient computation of Gaussian Markov random fields (GMRF) which can be used to speed up computations and provide fast approximations. GMRF are the key to providing good Gaussian approximations for the posterior marginals. INLA is based on providing Gaussian approximations to densities like

$$\pi(\mathbf{x}|\theta, \mathbf{y}) \propto \exp\left\{-\frac{1}{2}\mathbf{x}^T \mathbf{Q} \mathbf{x} + \sum_{i \in \mathcal{I}} \log(y_i|x_i, \theta)\right\}$$

where \mathbf{Q} is the precision matrix of the GMRF. Note that if \mathbf{Q} is a known matrix, its determinant (sometimes termed Jacobian) can be ignored at this stage as the posterior distribution can be rescaled later. This distribution may be subject to a set of linear constraints $\mathbf{Ax} = \mathbf{e}$. In any case, the approximation will result in a Gaussian distribution with mean \mathbf{x}^* and precision matrix $\mathbf{Q}^* = \mathbf{Q} + \text{diag}(\mathbf{c}^*)$ (see Rue et al. 2009, Sect. 2 for details). If linear constraints are present, the mean and precision matrices of the Gaussian approximation are conveniently corrected.

These constrained models are useful for fitting geostatistical models and adjacency-based spatial correlation effects for areal data (e.g., using an intrinsic conditional autoregressive model). Other spatial and temporal random effects can be modeled by using intrinsic GMRFs with linear constraints (see Rue and Held 2005, Chap. 3). Linear constraints are often employed to impose a sum-to-zero constraint on intrinsic GMRFs in order to make these effects identifiable. This is particularly important when dealing with complex spatiotemporal effects (Knorr-Held 2000).

71.2.2 Priors

So far, we have dealt with how the likelihood and the latent Gaussian Markov random fields are defined. As in all Bayesian approaches, a set of priors needs to be assigned to the parameters.

First of all, covariate coefficients in the linear predictor will be assigned a normal distribution with zero mean and precision τ . A similar distribution will be used for the random errors ε_i .

In principle, the latent random effects will be all Gaussian with zero mean. Hence, only the parameters in the precision matrix will need a prior. For the case in which the precision matrix is of the form $\tau \mathbf{Q}$, where \mathbf{Q} is a known matrix, τ can be assigned either a gamma, truncated normal, or improper flat (in the log-scale) prior. If the whole precision matrix is to be assigned a prior, then a Wishart distribution is available for correlated random effects of small dimension (up to 5). Finally, the INLA software provides other prior distributions. For example, correlation parameters, such as the ones used to model spatial autocorrelation, can be assigned a beta prior.

Note that, for simple models, these choices are equivalent to setting a conjugate prior distribution and that in all cases the prior parameters are supposed to be known (i.e., these cannot be assigned a prior in turn). It should be mentioned that these priors are the ones implemented in the INLA software (available from <http://www.r-inla.org>), but user-defined priors can be used as well by providing the mathematical expression for them.

Other priors can be built on upon simpler prior specifications. For example, spatially varying coefficients on a covariate can be implemented by using a prior which is the sum of independent and spatially correlated random effects. More information about how priors can be specified are available at <http://www.r-inla.org/models/priors>.

71.2.3 Model Criticism and Selection

INLA provides a number of ways of comparing and assessing models. First of all, an approximation to the marginal likelihood $\pi(\mathbf{y})$ is provided. This approximation is based on

$$\tilde{\pi}(\mathbf{y}) = \int \frac{\pi(\mathbf{x}, \theta, \mathbf{y})}{\pi_G(\mathbf{x}|\theta, \mathbf{y})} |_{x=x^*(\theta)} d\theta$$

where $\pi(\mathbf{x}, \theta, \mathbf{y}) = \pi(\theta)\pi(\mathbf{x}|\theta)\pi(\mathbf{y}|\mathbf{x}, \theta)$. Models with a larger value of the marginal likelihood will be preferred. Also, marginal likelihood can be used to compute Bayes factors in order to compare models.

Predictive measures can also be computed very easily. In particular, INLA can compute the predictive distribution of y_i given all the other observations, that is, $\pi(y_i|\mathbf{y}_{-i})$. Following Pettit (1990), INLA reports the probability integral transform (PIT):

$$PIT_i = Prob(y_i^{new} \leq y_i | \mathbf{y}_{-i})$$

This criterion has been used to assess the validity of spatial models in disease mapping and it avoids the use of other sampling-based methods which may be less accurate (Marshall and Spiegelhalter 2003).

Roos and Held (2011) discuss sensitivity to priors for binary data using the conditional predictive ordinate (CPO, Geisser 1993), which is defined as $\pi(y_i|\mathbf{y}_{-i})$. They use the mean logarithmic CPO to build the following statistic as a measure of the predictive quality of the model:

$$\overline{CPO} = -\frac{1}{n} \sum_i^n \log(\pi(y_i|\mathbf{y}_{-i}))$$

Lower values of \overline{CPO} indicate a better model. As the authors state, this criterion can easily be extended to other hierarchical models. Held et al. (2010) compare the

CPO and PIT between “exact” Bayesian inference (using MCMC) and approximate inference (with INLA) showing that the approximated values are very close in general to the exact ones.

Finally, INLA can also compute the deviance information criterion (DIC, Spiegelhalter et al. 2002) which is a popular way of comparing Bayesian hierarchical models. The DIC also computes a measure of the effective number of parameters which is a measure of the complexity of the model.

71.2.4 Implementation

Besides the original paper, the authors have released a software (called INLA) which implements all the techniques mentioned here. In addition, an interface for the R programming language (R Development Core Team 2011) can be downloaded (from <http://www.r-inla.org>) which makes the use of the software easier and is able to produce summary statistics and plots of the results.

71.2.5 Other Features

In addition to an easy to use interface, the INLA software provides some other features. The joint posterior distribution of the hyperparameters can be computed. In addition, it is possible to define several linear combinations of the latent effects so that their posterior marginals are computed. Furthermore, if several of these linear combinations are computed, the joint correlation matrix can be computed as well, and this can be used to approximate the joint posterior distribution.

71.3 Spatial Models

Spatial dependence can be modeled in different ways in Bayesian hierarchical models (Banerjee et al. 2004). Given that INLA focuses on latent Gaussian models and given that the latent effects are Gaussian, spatial correlation can be embedded in the precision matrix. Furthermore, because of the Markov properties of the latent field, these variance-covariance matrices are often very sparse. How these methods can be applied to the different areas of spatial statistics is discussed below.

71.3.1 Geoadditive Mixed-Effects Models

Geoadditive models appear when regression models on a set of covariates are combined with other types of random effects (Kammann and Wand 2003). A geoadditive model will be based on modeling the mean μ_i at each location i on the sum of a set of fixed and random effects:

$$\mu_i = \mu + \mathbf{z}_i\beta + u_i + v_i$$

where \mathbf{z}_i is a vector of covariates and β the associated coefficients. \mathbf{u} is a vector of spatially correlated random effects, while \mathbf{v} is a vector of independent random effects.

Note that this modeling can be done regardless of the likelihood employed for the data. In the case of a generalized linear model, a convenient link function will be used to transform the linear predictor accordingly.

Other nonparametric approaches can be implemented taking advantage of this approach. Kammann and Wand (2003) and Ruppert et al. (2003) show how penalized splines (P-splines) can be expressed as a mixed-effects model. Lee and Durbán (2009) describe how P-splines and a CAR model can be used to model spatial data. They develop an expression of these models as mixed-effects models. Although this is not a fully Bayesian approach, these models could be fitted with INLA using the following representation:

$$\mu = \mathbf{X}\beta + \mathbf{Z}\mathbf{u}$$

Here \mathbf{X} and \mathbf{Z} represent design matrices for the fixed and random effects which have a particular structure derived from the fact that this mixed model represents a P-spline (see Sect. 4.9 in Ruppert et al. 2003, for details). A fully Bayesian approach to P-splines can be found in Lang and Brezger (2004), and it is based on imposing a prior on the coefficients γ of a design matrix \mathbf{B} (based on the basis functions):

$$\mu = \mathbf{B}\gamma$$

Different priors on γ lead to different types of splines (Fahrmeir and Kneib 2011). For producing smoothed values of an observed covariate using P-splines, the prior should be a random walk. To achieve spatial smoothing, the prior on γ should be a GMRF with spatial structure. See Lang and Brezger (2004) for details on how to define \mathbf{B} and the prior of γ for spatial smoothing.

71.3.2 Disease Mapping

The analysis of public health data has played an important role in the development of spatial statistics in the last two decades. Besag et al. (1991) provided a suitable model in which spatial correlation and unstructured variation are combined in a geoadditive way which is also computationally appealing. Other authors have extended this model later, some of them for spatiotemporal disease mapping.

It should not be forgotten that disease mapping is a particular example of the analysis of lattice data. In this case, observations are aggregated over some region (counties, states, health districts, etc.) and spatial models assume that neighboring areas will have similar behavior. Here, dependence is between neighbors and a popular criterion is that two areas are neighbors if they share a common boundary.

Besag et al. (1991) proposed the use of two latent random effects: a spatially correlated one \mathbf{u} and an independent one \mathbf{v} . The first will account for any spatial correlation and the second will account for any other unstructured difference between the regions. While the nonstructured random effects are Gaussian with zero mean and precision τI_n (where I_n is the identity matrix of size $n \times n$), the spatially correlated random effects are defined using conditional distributions given the values at the neighbors. This is equivalent to using an intrinsic GMRF (Rue and Held 2005, Chap. 3), which is known as intrinsic conditionally autoregressive (CAR) model.

In order to encode this spatial information into a GMRF with zero mean and precision Q , we will make use of the Markov property to note that if areas i and j are independent given the remaining areas, then $Q_{ij} = Q_{ji} = 0$. Hence, the precision matrix Q will be very sparse, and the algorithms described in Rue and Held (2005) can be used for fast sampling from this GMRF.

In particular, the intrinsic CAR precision matrix is defined as

$$Q_{ij} = \kappa \begin{cases} n_i & i = j \\ -1 & i \sim j \\ 0 & \text{otherwise} \end{cases}$$

Here $i \sim j$ means that areas i and j are neighbors, κ is a conditional precision, and n_i is the number of neighbors of area i . This makes the conditional distribution of $u_i | \mathbf{u}_{-i}, \kappa$ Gaussian with mean $\frac{1}{n_i} \sum_{j \sim i} u_j$ and variance $\frac{1}{\kappa n_i}$.

Note that the intrinsic CAR is an improper GMRF of rank $n - 1$. For this reason the constraint $\sum_i u_i = 0$ is added so that these effects can be identified. This is a common assumption for random effects based on intrinsic GMRF (Martino and Rue 2010).

A proper version of the intrinsic CAR model is available and it has a precision matrix similar to the previous one but adding a term $d > 0$ to the diagonal elements, so that they become $Q_{ii} = n_i + d$. $\log(d)$ is assigned a log-gamma prior distribution by default. Note that the main point of this model is to make the precision matrix strictly diagonally dominant so that it becomes invertible and the prior distribution is a proper one.

A more general approach is obtained when the precision matrix is defined as

$$\mathbf{Q} = (I - \frac{\rho}{\lambda_{max}} \mathbf{C})$$

This can be used to define a general CAR spatial effect by taking \mathbf{C} as a matrix of spatial weights (see Chap. 9 in Bivand et al. 2008, to see how different spatial weights can be defined). ρ represents the spatial correlation (and it can be assigned a prior) and takes values between 0 and 1 because the weight matrix is \mathbf{C} divided by λ_{max} , its maximum eigenvalue, and by default a Gaussian prior is on $\log(\rho)$. Note that this will produce a proper distribution for the spatially correlated random effects. Negative spatial autocorrelation is often ignored in disease mapping.

In this general case, the conditional distribution of u_i is

$$u_i | \mathbf{u}_{-i}, \kappa \sim N\left(\rho \frac{\sum_{j \neq i} w_{ij} u_j}{w_{i+}}, \frac{1}{\kappa w_{i+}}\right)$$

where $w_{ij} = c_{ij}/\lambda_{\max}$ and $w_{i+} = \sum_{j=1}^n w_{ij}$. Note that if \mathbf{C} is row standardized, then $\lambda_{\max} = 1$ and $w_{i+} = 1$ and the marginal distribution has a simpler form.

71.3.3 Geostatistical Models

In addition to fitting a model to the data, geostatistics focuses on predicting a continuous surface (often approximated by a discrete grid of points) so these models are often computationally very expensive. Spatially correlated random effects are built for the set of sampling locations, which may lead to trouble if the number of locations is large. Geostatistical models are not restricted to Gaussian likelihoods, as described in Banerjee et al. (2004) and Diggle and Ribeiro (2007), and they can be used to model other types of data using a geostatistical latent effect.

Spatial correlation in geostatistical models is built upon the distances between the sampling points, usually using a decaying function on the distance. For example, a simple covariance function is defined such as $\Sigma_{ij} = \sigma^2 \exp(-d_{ij}/\varphi)$. Here d_{ij} is the distance between points i and j , and φ is a parameter to control for the spatial scale. Once the model is fitted, prediction relies on the posterior distributions of the parameters and the covariances for the points in the grid.

A more general class of spatial covariance is provided by the Matérn correlation function, of which the exponential decaying function is a particular example. The Matérn covariance is defined as

$$\Sigma_{ij} = \sigma^2 \frac{\tau^\kappa K(\tau, \kappa)}{2^{\kappa-1} \Gamma(\kappa)}; \tau = \alpha_\kappa d_{ij}/\varphi$$

$K(\cdot, \kappa)$ is the modified Bessel function of order κ and $\Gamma(\cdot)$ the gamma function. α_κ and φ can be used to control the scale of the spatial variation. Setting κ to 0.5 leads to an exponential covariance. Other values of κ will lead to other known spatial covariance functions (Eidsvik et al. 2009).

When it comes to provide a prediction on the grid, INLA treats the observation at each point on the grid as a missing value. This makes INLA compute the marginal posterior distribution at that point so that summary statistics can be obtained later.

In this approach, modeling and prediction occur on a regular grid, and observations need to match to some location in the grid. Lindgren et al. (2011) aim at modeling the geostatistical model by using a mesh based on a triangulation of the sampling points (instead of a regular grid) and stochastic partial differential equations (SPDE). In this approach, the spatially distributed effect \mathbf{u} is

$$u(s) = \sum_{k=1}^n \psi_k(s) w_k, s \in \mathbb{R}^2$$

where $\{\psi_k\}$ are some basis functions, $\{w_k\}$ are Gaussian distributed weights, and n is the number of points in the triangulation used to split the study area. As this is a more complex approach, the reader is referred to the original paper (Lindgren et al. 2011) and the gentle introduction by Cameletti et al. (2011) for details on how the basis functions and weights are taken.

Finally, INLA can be used for geostatistical design. Methods and results discussed in Diggle et al. (2010) for preferential sampling can be reproduced with INLA (see the Case Studies section in <http://www.r-inla.org>). Anisotropic models could also be employed, as discussed in Fuglstad (2011), and use of these models is being integrated into the software package.

71.3.4 Point Process Models

Rue et al. (2009) show an example of the analysis of a point pattern with INLA using a Poisson process. Rather than modeling the continuous intensity of the point process, they divide the study area in N disjoint cells (not necessarily of equal size) and model the data as coming from a counting process. Hence, the response variable y_i represents the number of occurrences of the process in square $w_i; i = 1, \dots, N$. For simplicity a square lattice may be employed. In a square lattice all the squares have the same area, and spatially correlated random effects can be defined similarly as in lattice data (i.e., two squares are neighbors if they have a common boundary).

In their example, Rue et al. (2009) use a hierarchical Poisson process to model the number of trees in each square using a log-Gaussian Cox process (LGP). In this case, the intensity function is $\lambda(s) = \exp\{Z(s)\}; s \in \mathbb{R}^2$, where $Z(s)$ is a Gaussian field at $s \in \mathbb{R}^2$.

Hence, y_i is the observed number of occurrences in cell w_i . If η_i is the realization of $Z(s_i)$, then $\pi(y|\eta) = \prod_i \pi(y_i|\eta_i)$, where $\pi(y_i|\eta_i)$ represents a Poisson distribution with mean $|w_i| \exp(\eta_i)$. $|w_i|$ is the area of cell w_i .

In turn, η_i is modeled according to a number of covariates plus some random effects:

$$\eta_i = X_i \beta + u_i + v_i$$

\mathbf{u} and \mathbf{v} are modeled in a similar way as with the lattice data case. v_i are independent Gaussian with zero mean and variance σ_v^2 so that they represent independent variation between the squares. On the other hand, u_i are modeled using a second-order polynomial intrinsic GMRF. In this way, first-, second-, and third-order neighbors

are taken into account, each one with a different weight, to mimic thin plate splines. See Rue and Held (2005) for details.

Simpson et al. (2011) extend the ideas in Lindgren et al. (2011) to model the latent LGP in a continuous way using a mesh on the study area. They show that this is a better approach that reduces the computational burden as a mesh is used instead of a regular grid and there is no need to aggregate cases into small cells.

More complex models cannot be fully addressed using INLA, in particular, those for which a closed likelihood does not exist as, for example, Gibbs processes. In a Gibbs process, future observations depend on present observations and, hence, producing a likelihood in closed form is not feasible.

71.4 Examples

As it happens, INLA is one of many alternatives for fitting Bayesian hierarchical models. In this section we provide a comparison to other software available for the R programming language, including computing times. Our aim here is not to provide a full comparison of computation times but to indicate how different approaches compare in terms of time and accuracy of results when used to fit a similar model to the same data set.

71.4.1 Geostatistics

For geostatistical models, we will use the Rongelap data set analyzed in several works on model-based geostatistics (Diggle and Ribeiro 2007). This data set records radionuclide concentration at 157 different locations, and the interest is on providing an estimate of the concentration over the whole Rongelap island.

As INLA makes computation on a regular grid, we have considered a 5×5 regular grid on one of the clusters in the northeast part of the island to make a fair comparison between computing times. We have used the INLA software (using the Laplace approximation) and the R package geoRglm, which provides model fitting using MCMC. The different computation times are shown in Table 71.1, while a map comparing the different estimates is shown in Fig. 71.1.

71.4.2 Lattice Data

For the case of lattice data, we have used the number of total malignant neoplasms mortalities in Georgia in 1999. We have fitted the model proposed in Besag et al. (1991) with population density as a covariate. In this case, we have used the INLA software as well as WinBUGS. Times are available in Table 71.1 and a graphical comparison of the estimates is available in Fig. 71.2.

Table 71.1 Summary of computation times for different problems, softwares, and fitting methods

Software	Method	Geostatistics		Lattice data		Point patterns	
		# Iter.	Time (s)	# Iter.	Time (s)	# Iter.	Time (s)
R-INLA	INLA	—	0.251	—	0.422	—	0.758
geoRglm	MCMC	22,000	0.409	—	—	—	—
WinBUGS	MCMC	—	—	22,000	11.420	22,000	35.336

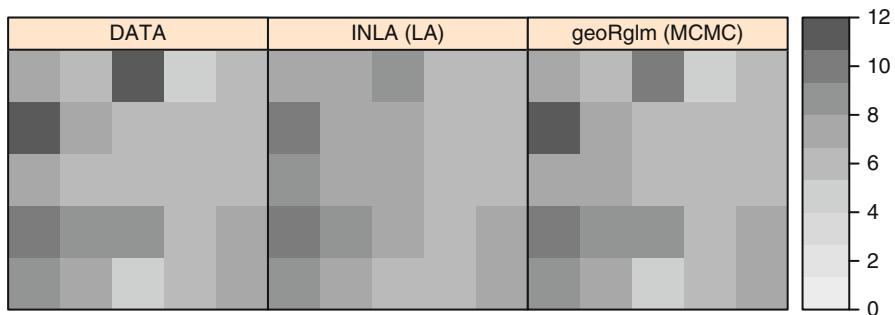


Fig. 71.1 Estimates of the radionuclide concentration using different methods: Integrated nested Laplace approximation (INLA) and MCMC (using geoRglm)

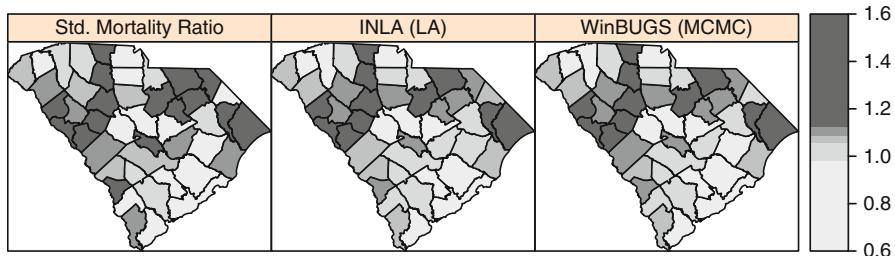


Fig. 71.2 Estimates of the relative risk using different methods: Standardized mortality ratio (SMR), integrated nested Laplace approximation (INLA), and MCMC (using WinBUGS)

71.4.3 Point Patterns

Finally, a point pattern has been included; we have performed an analysis of the Japanese pines data set available in R package Spatstat. This data set provides the location of Japanese pine saplings in a square region in a natural forest. Again, model fitting with INLA requires the use of a regular square grid so that the data are the number of saplings in each grid square. A 10×10 square grid has been used in this case, and the model to account for spatial dependence is the same as in the

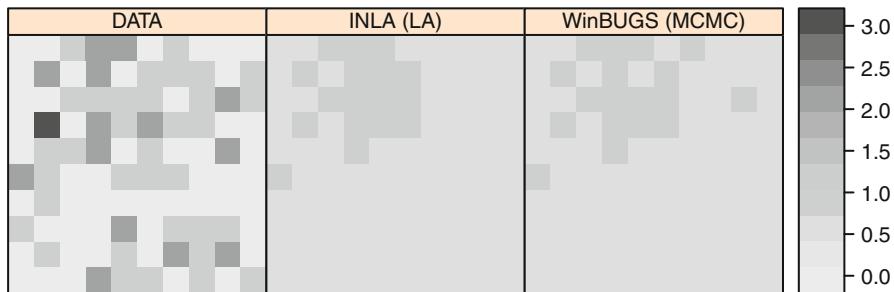


Fig. 71.3 Estimates of the number of saplings per square using two different methods: Integrated nested Laplace approximation (INLA) and MCMC (using WinBUGS)

previous example (Besag et al. 1991). This will also give us an idea of how INLA behaves as the grid size increases.

Figure 71.3 summarizes the fitted number of saplings and computing times are available in Table 71.1. It is worth noting how the differences between INLA and WinBUGS have increased now.

71.5 Conclusions

The integrated nested Laplace approximation developed in Rue et al. (2009) provides a series of approximations for the posterior marginals of the parameters of a Bayesian hierarchical model in which the latent effects are a Gaussian Markov random field. This family of models covers a good number of Bayesian hierarchical models, including several of those most used in spatial statistics. In addition, Markov properties are very convenient in dealing with spatial data and they can be used to model local dependence. Besides an approximation to the posterior marginals of the parameters in the model, INLA can compute several criteria for model criticism and selection, such as PIT and the DIC.

Regarding spatial models, INLA has been used to tackle problems in the analysis of lattice data, geostatistics, and point processes. In all cases, spatial dependence is modeled via the precision matrix of Gaussian random effects. The recent developments by Lindgren et al. (2011) allow for continuous modeling of latent spatial effects, which avoids the use of a grid and provides a good computational approach as well.

The availability of associated software that implements all these methods provides a suitable framework for their wider use. Other external software may be required to display the results in maps or create adjacency matrices for the analysis of lattice data. For this reason, the authors of the INLA software have provided an interface to the R programming language. The R-INLA web site (<http://www.r-inla.org>) provides the latest version of the software and its documentation as well as an updated list of published and working papers.

Acknowledgments Virgilio Gómez-Rubio has been supported by the Spanish Ministry of Science and Innovation (project MTM 2008–03085) and Junta de Comunidades de Castilla-La Mancha (project PPIC11-0183-7474).

References

- Banerjee S, Gelfand AE, Carlin BP (2004) Hierarchical modeling and analysis for spatial data. Chapman and Hall/CRC, Boca Raton
- Besag J, York J, Mollie A (1991) Bayesian image restoration, with two applications in spatial statistics. *Ann Inst Stat Math* 43(1):1–59
- Bivand RS, Pebesma EJ, Gómez-Rubio V (2008) Applied spatial data analysis with R. Springer, New York
- Cameletti M, Lindgren F, Simpson D, Rue H (2012) Spatio-temporal modeling of particulate matter concentration through the SPDE approach. *Adv Statistical Anal.* <http://dx.doi.org/10.1007/s10182-012-0196-3>
- Diggle P, Ribeiro PJ (2007) Model-based geostatistics. Springer, New York
- Diggle PJ, Menezes R, TI S (2010) Geostatistical inference under preferential sampling. *J R Stat Soc Ser C Appl Stat* 59(2):191–232
- Eidsvik J, Martino S, Rue H (2009) Approximate Bayesian inference in spatial generalized linear mixed models. *Scand J Stat* 36(1):1–22
- Fahrmeir L, Kneib T (2011) Bayesian smoothing and regression for longitudinal, spatial and event history data. Oxford University Press, New York
- Fuglstad GA (2011) Spatial modelling and inference with SPDE-based GMRFs. Master's thesis, Norwegian University of Science and Technology, Norway
- Geisser S (1993) Predictive inference: an introduction. Chapman & Hall, New York
- Gelman A, Carlin JB, Stern HS, Rubin DB (2003) Bayesian data analysis, 2nd edn. Chapman and Hall/CRC, Boca Raton
- Held L, Schödle B, Rue H (2010) Posterior and cross-validatory predictive checks: a comparison of MCMC and INLA. In: Kneib T, Tutz G (eds) Statistical modelling and regression structures – Festschrift in Honour of Ludwig Fahrmeir. Springer, Berlin, pp 91–110
- Illian JB, Martino S, Sorbye S, Gallego-Fernandez J, Travis J (2012) Fitting complex ecological point processes with integrated nested Laplace approximation (inla). *Methods Ecol Evol*. <http://www.methodsinecologyandevolution.org/view/0/accepted.html>
- Kammann EE, Wand MP (2003) Geoadditive models. *J R Stat Soc Ser C Appl Stat* 52(1):1–18
- Knorr-Held L (2000) Bayesian modelling of inseparable space-time variation in disease risk. *Stat Med* 19:2555–2567
- Lang S, Brezger A (2004) Bayesian p-splines. *J Comput Graph Stat* 13(1):183–212
- Lee DJ, Durbán M (2009) Smooth-car mixed models for spatial count data. *Comput Stat Data Anal* 53(8):2968–2979
- Lindgren F, Rue H, Lindström J (2011) An explicit link between Gaussian fields and Gaussian Markov random fields: the SPDE approach (with discussion). *J R Stat Soc Ser B* 73(4):423–498
- Marshall EC, Spiegelhalter DJ (2003) Approximate cross-validatory predictive checks in disease mapping models. *Stat Med* 22(10):1649–1660
- Martino S, Rue H (2010) Case studies in Bayesian computation using INLA. In: Mantovan P, Secchi P (eds) Complex data modeling and computationally intensive statistical methods, Contributions to statistics. Springer, New York, pp 99–114
- Pettit LI (1990) The conditional predictive ordinate for the normal distribution. *J R Stat Soc Ser B Methodol* 52(1):175–184
- R Development Core Team (2011) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna. <http://www.R-project.org/>. ISBN 3-900051-07-0

- Roos M, Held L (2011) Sensitivity analysis in Bayesian generalized linear mixed models for binary data. *Bayesian Anal* 6(2):259–278
- Rue H, Held L (2005) Gaussian Markov random fields. Theory and applications. Chapman & Hall, New York
- Rue H, Martino S (2007) Approximate Bayesian inference for hierarchical Gaussian Markov random field models. *J Stat Plan Inference* 137(10, SI):3177–3192
- Rue H, Martino S, Chopin N (2009) Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *J R Stat Soc Ser B Stat Methodol* 71(Pt 2):319–392
- Ruppert D, Wand MP, Carroll RJ (2003) Semiparametric regression. Cambridge University Press, New York
- Schroedle B, Held L, Riebler A, Danuser J (2011) Using integrated nested Laplace approximations for the evaluation of veterinary surveillance data from Switzerland: a case-study. *J R Stat Soc Ser C Appl Stat* 60(Pt 2):261–279
- Simpson D, Illian J, Lindgren F, Sørbye SH, Rue H (2011) Going off grid: computationally efficient inference for log-Gaussian Cox processes. Preprint Statistics 10/2011. Norwegian University of Science and Technology, Trondheim
- Spiegelhalter DJ, Best NG, Carlin BP, Van der Linde A (2002) Bayesian measures of model complexity and fit (with discussion). *J R Stat Soc Ser B* 64(4):583–616

Peter Congdon

Contents

72.1	Introduction	1419
72.2	Spatially Autoregressive Regression in Spatial Econometrics	1420
72.3	Discrete Outcomes: Conditional Priors in Spatial Epidemiology	1423
72.4	Spatial Covariation in Continuous Space	1427
72.5	Space-Time Models	1429
72.6	Focused Clustering Models	1431
72.7	Conclusions	1432
	References	1433

Abstract

Spatial statistics has in the last decade or two emerged as a major sub-specialism within statistics. Applications areas are diverse, and there is cross-fertilization with methodologies in other disciplines (econometrics, epidemiology, geography, geology, climatology, ecology, etc). This chapter reviews three major settings and techniques that have attracted attention from statisticians: spatial econometrics and simultaneous autoregressive models, spatial epidemiology and conditional autoregressive models, and geostatistical methods for point pattern data. The review is oriented to Bayesian inferences for such models, including discussion of choice of prior densities, questions of identification, outcomes of interest, and methods of estimation (using Markov chain Monte Carlo).

P. Congdon

School of Geography, Queen Mary University of London, London, UK

e-mail: p.congdon@qmul.ac.uk

72.1 Introduction

Bayesian applications in spatial statistics have multiplied considerably in the last two decades, facilitated by improved estimation using Markov chain Monte Carlo (MCMC) methods and by advances in relevant statistical theory. Application areas where Bayesian ideas have impacted include spatial epidemiology, spatial ecology, spatial econometrics and political science, and geostatistics. In spatial epidemiology, Bayesian studies include spatial smoothing of rare health outcomes, modelling spatial clustering in disease risks (e.g., Richardson et al. 2004), and models for health impacts of environmental point sources (Wakefield and Morris 2001), while in spatial ecology, applications include habitat and remote sensing models (Carroll et al. 2010). Applications in spatial econometrics concentrate on models for behavior by economic actors (house purchasers, firms, etc) involved in spatially defined behaviors (e.g., LeSage and Pace 2009), while spatial applications in political science (Beck et al. 2006) focus on spatially defined electoral and legislative processes. Another major application context is the continuous spatial framework of geostatistics with diverse applications including geology, infectious epidemiology, and meteorology (e.g., Ecker and Gelfand 1997; Diggle and Ribeiro 2007; Schur et al. 2011).

Bayesian analysis in such applications is distinct from frequentist approaches in the need to consider the specification of prior densities for parameters θ . Such densities can potentially summarize existing evidence (e.g., from previous studies) where available or may express subject matter based constraints, such as confining a spatial correlation parameter to positive values. Prior densities may vary in their informativeness, meaning essentially the degree of concentration in the mass: a diffuse or flat prior will spread the prior density over a wide range of values (e.g., as in a uniform prior for a probability or rate), whereas an informative prior will concentrate potential values within a narrower range. The prior density $\pi(\theta)$ for a parameter is updated by the likelihood of the data $L(\theta|y) = p(y|\theta)$, and posterior inferences are based on the updated parameter density $\pi(\theta|y) = kp(y|\theta)\pi(\theta)$. Typically modern spatial data analysis using Bayesian principles will also use Markov chain Monte Carlo (MCMC) sampling methods in the updating stage, and there may be advantages in being able to use particular MCMC sampling methods such as Gibbs sampling (Casella and George 1992), which involves repeated sampling from the full conditional density of a parameter. Choice of prior density may be important in facilitating MCMC sampling, as illustrated in some of the techniques described below. As well as facilitating estimation of complex spatial models, MCMC techniques aid in related inferences: examples include posterior probabilities of elevated disease risk, also called exceedance probabilities (Richardson et al. 2004; Hossain and Lawson 2006).

72.2 Spatially Autoregressive Regression in Spatial Econometrics

Consider the normal linear regression with continuous outcomes y_i , predictor vector X_i , and errors e_i , $i = 1, \dots, n$. In applications to observations for discrete spatial

units (also called lattice data), an *iid* assumption regarding the errors is likely to be invalid, and instead there will often be covariation in errors for closely co-located areas i and j . For example, if the model is for crime rates, then positive regression residuals may tend to be spatially clustered because crime rates themselves are spatially clustered (e.g., Ratcliffe 2010). In spatial econometrics, an adaptation of linear regression tackles potential spatial correlation in terms of spatially lagged dependence in errors or observations. This is analogous to similar forms of lagged dependence often applied in time series regression, such as first- and possibly higher-order lags in the dependent variable, and serially correlated errors in time.

Thus, consider an $n \times n$ matrix C of contiguity dummies, with $c_{ij} = 1$ if areas i and j are adjacent, and $c_{ij} = 0$ otherwise (with $c_{ii} = 0$). Alternatively distance-based interactions might be specified, for instance, $c_{ij} = \exp(-\gamma d_{ij})$, where $\gamma > 0$ reflects distance decay. From C may be obtained the row-standardized matrix $W = [w_{ij}] = [c_{ij} / \sum c_{ij}]$. The most general model, known as a spatial autoregressive regression, includes a spatial lag in both errors and observations, namely (for y of dimension $n \times 1$, predictors X of dimension $n \times p$, and β of dimension $p \times 1$),

$$\begin{aligned} y &= \lambda W_1 y + X\beta + \varepsilon \\ \varepsilon &= \rho W_2 \varepsilon + u \\ u &\sim N(0, \sigma^2 I) \end{aligned}$$

where λ and ρ are unknown correlation parameters. The coefficients λ and ρ have bounds $\left\{ \frac{1}{\omega_{1,\min}}, \frac{1}{\omega_{1,\max}} \right\}$ and $\left\{ \frac{1}{\omega_{2,\min}}, \frac{1}{\omega_{2,\max}} \right\}$, respectively, where $\omega_{j,\min}$ and $\omega_{j,\max}$ are the minimum and maximum eigenvalues of W_j . For W standardized within rows, ω_{\max} is 1, and since spatial correlation is usually positive, a prior on λ or ρ constrained to $[0, 1]$ is often used, for example, a beta prior, $\pi(\lambda) = Beta(a_\lambda, b_\lambda)$ (LeSage and Pace 2009, p. 142).

A widely used scheme for error dependence is a reduced version of the above model, known commonly as the spatial error model, with $W_1 = 0$. This model expresses spatial covariation in errors caused by omitted predictor variables, measurement errors, and possible mismatch between the spatial units used, and the scale at which the process occurs (Anselin and Bera 1998). The corresponding likelihood is

$$L(\beta, \sigma^2, \rho | y) = (2\pi)^{-n/2} \sigma^{-n} |I - \rho W| \exp\left(-\frac{1}{2\sigma^2} \varepsilon' \varepsilon\right)$$

where $\varepsilon = (I - \rho W)(y - X\beta)$. Spatially lagged effects in the dependent variable rather than errors (i.e., $W_2 = 0$ in the general model above) lead to the spatial autoregressive model

$$y = \lambda W y + X\beta + u$$

where u is an *iid* error. This model is often used to represent neighborhood diffusion or spillover effects, as in applications to technical innovation and house prices, respectively. The corresponding likelihood is

$$L(\beta, \sigma^2, \lambda|y) = (2\pi)^{-n/2} \sigma^{-n} |I - \lambda W| \exp\left(-\frac{1}{2\sigma^2} u'u\right)$$

where $u = (I - \lambda W)(y - X\beta)$.

To demonstrate the derivation of conditional posterior densities and appropriate forms of MCMC sampling, consider the spatial autoregressive model (without spatially correlated errors). One may assume prior independence between λ and the other parameters, but assume a normal prior for β that conditions on the sampled value of σ^2 . Thus, with $IG(a, b)$ denoting the inverse Gaussian density, one has

$$\begin{aligned}\pi(\sigma^2) &= IG(\sigma^2|a_0, b_0) = \frac{b_0^{a_0}}{\Gamma(a_0)} (\sigma^2)^{-(a_0+1)} \exp(-b_0/\sigma^2) \\ \pi(\beta|\sigma^2, c_0, d_0) &= N(c_0, d_0\sigma^2)\end{aligned}$$

where c_0 is a vector and d_0 is a matrix. The prior on λ is uniform between bounds determined by the eigenvalues ω of W :

$$\pi(\lambda|\omega, W) = U\left(\frac{1}{\omega_{\min}}, \frac{1}{\omega_{\max}}\right)$$

The combination of priors and likelihood assumptions determines the form of conditional density for each parameter θ_r , namely, that part of $L(\theta|y)\pi(\theta)$ varying in θ_r . Choice of MCMC sampling depends on the form of density $p(\theta_r|\theta_{[r]}, y)$ for a particular parameter θ_r conditional on all other parameters $\theta_{[r]}$. To implement Gibbs sampling usually requires that these full conditional densities have a known form that permits direct sampling. Letting $A = I - \lambda W$, the full conditional density for β has a normal form, permitting Gibbs sampling, namely,

$$\begin{aligned}p(\beta|\sigma^2, \rho, y) &= N(c_1, d_1\sigma^2) \\ c_1 &= d_1(X'Ay + d_0^{-1}c_0) \\ d_1 &= (X'X + d_0^{-1})^{-1}\end{aligned}$$

The full conditional density for σ^2 is also inverse Gaussian (again allowing Gibbs sampling) with form

$$\begin{aligned}p(\sigma^2|\beta, \rho, y) &= IG(a_1, b_1) \\ a_1 &= a_0 + 0.5n, \quad b_1 = b_0 + 0.5(Ay - X\beta)'(Ay - X\beta)\end{aligned}$$

However, the full conditional density for λ has the form

$$p(\lambda|\sigma^2, \beta, y) = k|I - \lambda W| \exp\left(-\frac{1}{2\sigma^2} u'u\right) U\left(\frac{1}{\omega_{\min}}, \frac{1}{\omega_{\max}}\right)$$

where k is an unknown constant. This is not a standard density, and so more general Metropolis or Metropolis-Hastings sampling is needed. Let $p(\lambda^{(t)})$ denote the value of $p(\lambda|\sigma^2, \beta, y)$ at the current value $\lambda^{(t)}$ in an MCMC sampling sequence $t = 1, \dots, T$. Let $p(\lambda_{new})$ be the value of the same conditional density at a candidate value generated by a proposal density. Let h be a random number between 0 and 1. Then, in Metropolis sampling, the candidate value replaces the current value either if $p(\lambda_{new}) > p(\lambda^{(t)})$ or if $p(\lambda_{new}) < p(\lambda^{(t)})$ but $h < p(\lambda_{new})/p(\lambda^{(t)})$. An alternative to a uniform or beta prior on λ (or ρ) mentioned by LeSage and Pace (2009, p. 139) is a prior defined over a grid of feasible values $\{\lambda_1, \dots, \lambda_L\}$, usually with an equal prior probability on each value λ_l . This allows pre-calculation of the log determinants of $I - \lambda_l W$, so lessening the computational burden during MCMC sampling.

While developed for continuous data, these techniques can be adapted to binary, multinomial, or ordinal outcomes using latent outcome representations. For binary data defined over areas $i = 1, \dots, n$

$$y_i \sim Bern(\pi_i)$$

the spatial autoregressive and spatial error models can be applied using a latent variable model, sometimes denoted the spatial probit model, whereby

$$\begin{aligned} y_i = 1 &\quad \text{if} & z_i > 0 \\ y_i = 0 &\quad \text{if} & z_i \leq 0 \end{aligned}$$

where z_i can be interpreted as the utility difference $U_{1i} - U_{0i}$ between binary options, with $Pr(y_i = 1) = Pr(U_{1i} > U_{0i}) = Pr(z_i > 0)$ (Smith and LeSage 2004). For example, the spatial autoregressive model for dichotomous outcomes based on the latent variable representation is

$$z = \lambda Wz + X\beta + u, \quad u \sim N(0, I)$$

so that

$$\begin{aligned} z &= (I - \lambda W)^{-1} X\beta + v \\ v &= (I - \lambda W)^{-1} u \sim N_n(0, [(I - \lambda W)'(I - \lambda W)]^{-1}) \end{aligned}$$

The variance of the residuals is preset for identifiability, while priors on λ and β follow schemes such as those discussed above.

72.3 Discrete Outcomes: Conditional Priors in Spatial Epidemiology

Simultaneous autoregressive schemes are primarily designed for continuous univariate responses, whereas count variables (usually leading to Poisson or binomial likelihoods) are common in health and ecological applications. Although transformations of count variables may be applied, leading to approximate normality (e.g., the Anscombe transform), inverse transformation is sometimes subject to bias, and direct analysis of untransformed counts may be easier for multivariate outcomes. Also while simplifications using grid priors and pre-calculated determinants can be used, Bayesian estimation of simultaneous regression models may be burdensome in large datasets, requiring sampling from a high-dimension multivariate normal density, and inverse or determinant calculations for large matrices.

By contrast, conditional autoregressive priors are an alternative, especially for discrete data outcomes, provided they are consistent with valid joint priors. Instead of focusing on the joint multivariate distribution of the entire vector ε , conditional priors involve the univariate density of each area's error, ε_i , conditioning on errors in all other areas $\varepsilon_{[i]} = \{\varepsilon_j, j \neq i\}$. Certain restrictions on the form of the spatial weight matrix C and the conditional precision of the ε_i need to be followed to ensure a valid joint density is obtained from the collection of conditional priors (Besag and Kooperberg 1995).

Conditional priors can be used in all forms of generalized linear model, including linear regression with $y = X\beta + \varepsilon$. Under the conditional autoregressive or $CAR(\rho)$ prior (Bell and Broemeling 2000), one has

$$\varepsilon_i | \varepsilon_{[i]} \sim N\left(\rho \sum_{j \neq i} c_{ij} \varepsilon_j, \sigma^2\right)$$

where the conditional mean is a weighted average of errors in other areas and ρ is bounded by the inverses of the minimum and maximum eigenvalues of C . Using a standardized weight matrix leads to what are often termed intrinsic conditional autoregressive or $ICAR(\rho)$ priors (Stern and Cressie 2000), with

$$\varepsilon_i | \varepsilon_{[i]} \sim N\left(\rho \sum_{j \neq i} c_{ij} \varepsilon_j / \sum_{j \neq i} c_{ij}, \sigma^2 / \sum_{j \neq i} c_{ij}\right)$$

The upper bound for ρ is now 1, and a uniform prior on ρ with values between 0 and 1 is often reasonable. If ρ is an unknown, then a common practice is to discretize the prior to equally spaced points (e.g., from 0.001 in spaces of 0.001 up to 0.999) to facilitate MCMC sampling.

A popular scheme, analogous to random walk priors in time series, in fact assumes $\rho = 1$. Additionally estimation of distance-decay parameters can be avoided by taking $c_{ij} = 1$ for adjacent areas, and $c_{ij} = 0$ otherwise. Define

$M_i = \sum_{j \neq i} c_{ij}$ as the number of areas adjacent to area i , and let L_i denote this collection of areas. Then, the *ICAR(1)* prior is

$$\varepsilon_i | \varepsilon_{[i]} \sim N\left(\sum_{j \in L_i} \varepsilon_j / M_i, \sigma^2 / M_i\right)$$

The joint prior version of this scheme is technically improper (Sun et al. 1999), but propriety is achieved in practice by recentering the sampled ε_i to sum to zero under MCMC sampling (Rodrigues and Assuncao 2008).

For example, suppose y_i denotes small-area disease counts, with expected events E_i obtained using region-wide incidence rates. The outcomes may be taken as Poisson, $y_i \sim Po(E_i \theta_i)$, where θ_i denotes relative risk of disease in area i . A classical approach (widely applied in area profiles of health outcomes) takes the θ_i as fixed effects, with (implicit) flat priors, and produces relative risk estimates $\hat{\theta}_i = y_i / E_i$. These may be misleading as indicators of varying disease patterns, since the resulting maps may be distorted by imprecisely estimated rates derived from low event counts or populations, and small changes in event totals may produce major shifts in estimates $\hat{\theta}_i$. Instead one possible plausible scheme for spatial borrowing of strength suggests two forms of underlying random variation: a smooth spatial signal ε_i following an *ICAR(1)* prior with variance σ_ε^2 and a *iid* term u_i for representing idiosyncratic local effects (Mollie 1996), leading to the so-called convolution prior with

$$\log(\theta_i) = X_i \beta + u_i + \varepsilon_i$$

where the *iid* errors are normal, $u_i \sim N(0, \sigma_u^2)$. Only the total error $t_i = u_i + \varepsilon_i$ is identified by the data, and estimates of variances σ_ε^2 and σ_u^2 may be sensitive to priors adopted (e.g., Yan 2006). Lee (2011) shows limitations of the convolution prior in both weak and strong spatial correlation situations.

To demonstrate the MCMC sampling involved in this model, define $\bar{\varepsilon}_i = \sum_{j \in L_i} \varepsilon_j / M_i$, and respecify the model for the log-relative risk as

$$\begin{aligned}\log(\theta_i) &= u_i + \varepsilon_i \\ u_i &\sim N(X_i \beta, \sigma_u^2)\end{aligned}$$

Then, the full conditional for each spatial error is

$$p(\varepsilon_i | \varepsilon_{[i]}, \sigma_\varepsilon^2, \beta, u_i) = k_1 \exp\{y_i \varepsilon_i - E_i \theta_i - 0.5 M_i (\varepsilon_i - \bar{\varepsilon}_i) / \sigma_\varepsilon^2\}$$

while the full conditional for each *iid* error is

$$p(u_i | \beta, \sigma_u^2, \beta, \varepsilon_i) = k_2 \exp\{y_i u_i - E_i \theta_i - 0.5 (u_i - X_i \beta) / \sigma_u^2\}$$

These conditionals can be sampled one at a time, or via block updating, using Metropolis-Hastings algorithms (Lee 2011).

Conventional spatial priors may not adequately model spatial discontinuities. Among ways to better represent discontinuities, and also avoid distorting the smooth spatial signal, the convolution prior may use a Student's t-distribution for the *iid* effect. This may be implemented using scale mixing, namely,

$$\begin{aligned} u_i &\sim N(0, \sigma_u^2 / \kappa_i) \\ \kappa_i &\sim G\left(\frac{v}{2}, \frac{v}{2}\right) \end{aligned}$$

where v is a degrees of freedom parameter. Areas with significantly lower κ_i are potential spatial outliers. The spatial prior itself may be adapted to be heavier-tailed: more robust alternatives including the double exponential prior

$$p(\varepsilon) \propto \chi \exp[-0.5\chi|\varepsilon_i - \varepsilon_j|^2]$$

where χ is a scaling parameter. Another option is mixture priors, such as

$$\log(\theta_i) = \gamma + u_i + \eta_i \varepsilon_{1i} + (1 - \eta_i) \varepsilon_{2i}$$

where ε_{1i} is *ICAR*(1), ε_{2i} follows the double exponential form, and η_i has a beta distribution (Lawson and Clark 2002).

Extending conditional autoregressive schemes to model multivariate spatial effects is relatively straightforward. Suppose there are J sets of spatial effects ε_{ji} for each area i . These might be relevant when there are J outcome variables, each with a spatially distributed regression residual, but can also be used in other ways: for example, in discrete mixtures over spatial effects or when regression coefficients show a spatial patterning. The latter scenario is sometimes denoted as spatially varying coefficient or SVC modelling and has the same intention in terms of representing spatial heterogeneity as techniques such as geographically weighted regression (Wheeler and Waller 2009). Multivariate CAR priors can also be applied in multilevel models; for example, a random intercept-random slope model with areas at level 2 would lead to a bivariate CAR prior.

Under the intrinsic multivariate conditional autoregressive or *IMCAR*(ρ, J) prior (Mardia 1988), the conditional prior for the i th area effect vector $\varepsilon_i = (\varepsilon_{1i}, \varepsilon_{2i}, \dots, \varepsilon_{Ji})$, given such effects for other areas, $\varepsilon_{[i]} = (\varepsilon_1, \dots, \varepsilon_{i-1}, \varepsilon_{i+1}, \dots, \varepsilon_n)$, is multivariate normal of dimension J with conditional outcome-specific means

$$\mu_{ji} = E(\varepsilon_{ji} | \varepsilon_{[i]}) = \rho \sum_{k \neq i} c_{ik} \varepsilon_{jk} / \sum_{k \neq i} c_{ik}$$

where ρ applies across all outcomes. When the c_{ik} are binary and based on contiguity, the outcome-specific conditional means are

$$\mu_{ji} = \rho \sum_{k \in L_i} \varepsilon_{jk} / M_i$$

namely, locality averages of spatial effects for outcome j , with corresponding within area conditional precision matrices

$$Prec(\varepsilon_i | \varepsilon_{[i]}) = M_i \Phi$$

where Φ is $J \times J$. Taking $\rho = 1$ in the $IMCAR(\rho, J)$ prior leads to the multivariate version of the $ICAR(1)$ prior.

72.4 Spatial Covariation in Continuous Space

The preceding discussion and examples consider continuous and discrete outcomes for zones (also called “lattice” data). Alternatively spatial data may consist of point data (e.g., geolocations for mineral deposits or for disease cases), sometimes denoted point pattern data, or aggregate data identified by grid-referenced location (Goovaerts and Gebreab 2008). For such data, the influence of interpoint or inter-location proximity on covariation in the outcome or regression errors needs to be explicitly considered or estimated.

Consider point or locational observations y_i at sites s_i in two-dimensional space, $s_i = (s_{1i}, s_{2i})$, with s_{1i} denoting longitude and s_{2i} denoting latitude. A starting point for estimating the effect of proximity is provided by a distance metric such as Euclidean interpoint distances, $d_{ij} = |s_i - s_j|$. A baseline assumption is that the spatial covariance matrix is isotropic, namely, independent of location and a function only of distance: so for points s and s' , separated by distance $d = |s - s'|$, one has $\Sigma(s, s') = \Sigma(d)$. Let $Y(s)$ and $\varepsilon(s)$ be $n \times 1$, with a predictor matrix $X(s)$ of dimension $n \times P$. Then,

$$\begin{aligned} Y(s) &= X(s)\beta + \varepsilon(s) \\ \varepsilon(s) &\sim N(0, \Sigma(d)) \end{aligned}$$

with $n \times n$ covariance matrix $\Sigma(d)$. Techniques such as variogram analysis can be used to explore covariation in regression residuals or investigate relevant assumptions such as isotropy (Irvine et al. 2007). Parametric functions can then be applied to represent $\Sigma(d)$.

Thus, consider $\Sigma(d) = \sigma^2 R(d)$ in terms of an overall variance σ^2 (defined along the diagonal when $i = j$ and $d_{ii} = 0$), and $R(d) = [r_{ij}(d_{ij})]$ reflecting correlations between the errors $\varepsilon(s_i)$ and $\varepsilon(s_j)$, usually such that $r_{ii}(0) = 1$ and $R(d)$ is positive definite (Diggle and Ribeiro 2007). Commonly used schemes include the exponential model

$$r_{ij} = \exp(-d_{ij}/\phi)$$

where ϕ is the range (distance at which spatial correlation ceases to be important) or the Gaussian function

$$r_{ij} = \exp(-d_{ij}^2/\phi^2)$$

In some cases there will be further *iid* variability (e.g., due to measurement error), leading to

$$\begin{aligned} Y(s) &= X(s)\beta + \varepsilon(s) + u \\ \Sigma(d) &= \sigma^2 R(d) + \tau^2 I \end{aligned}$$

where τ^2 is commonly known as the nugget variance, with the limiting variance as d_{ij} tends to zero being $\tau^2 + \sigma^2$ instead of σ^2 . Writing $V = \Sigma(d) = \sigma^2 R(d) + \tau^2 I$, $y = y(s)$, $X = X(s)$, the log-likelihood kernel is

$$-0.5 \log |V| - 0.5(y - X\beta)' V^{-1} (y - X\beta)$$

Prediction of y_{new} at a new site s_{new} under the linear model involves a vector of covariances $\lambda_i = Cov(s_{new}, s_i)$ between the new point and the sampled sites $s_i, i = 1, \dots, n$, and the prediction is then a weighted combination of the existing point values with weights w_i determined by

$$w = \lambda V^{-1}$$

Bayesian inference and estimation for such models may provide additional scope for inferences not possible under classical estimation approaches (e.g., ML or REML). An example is provided by Irvine et al. (2007), regarding the “effective range” or distance beyond which the correlation between observations, $\rho(d) = \Sigma(d)/\Sigma(0)$, is less than or equal to 0.05. On the other hand, computation may be slowed by MCMC calculations for high-dimension covariance matrices. An alternative is lower dimension spatial kernel methods such as discrete convolution priors (Higdon 2007). Precise estimation in such models may be facilitated by informative priors, for example, on distance-decay parameters such as ϕ or on the nugget-to-sill ratio $v^2 = \tau^2/\sigma^2$ in a reparameterized covariance matrix $\Sigma = \sigma^2(R + v^2 I)$ (Diggle and Ribeiro 2007, Chap. 7). Univariate or bivariate grid priors at selected points within a feasible range for ϕ and/or v^2 allow prior calculation of Σ or Σ^{-1} at the grid points and thus reduced computation.

Spatial covariance models can be defined for Poisson or binomial data (Diggle et al. 1998). Consider counts $y_i(s)$ assumed Poisson, with means $v(s_i)$. For equidispersed data a log link regression would then include a spatial error

$$\log(v(s_i)) = X(s_i)\beta + \varepsilon(s_i)$$

with $\varepsilon(s) \sim N_n(0, \sigma^2 R(d))$. However, for overdispersed data, both spatial and *iid* errors may be relevant, namely,

$$\log(v(s_i)) = X(s_i)\beta + \varepsilon(s_i) + u_i$$

with $u_i \sim N(0, \tau^2)$

72.5 Space-Time Models

Longitudinal spatial observations raise similar issues to those for panel data generally, such as the modelling of temporal autocorrelation and permanent area effects. The main strands in spatial modelling (spatial econometrics, spatial epidemiology, and geostatistics) all have space-time representations, though Bayesian modelling in some applications has been relatively limited. Thus spatio-temporal variations on spatial lag and spatial error models have only recently been considered in Bayesian terms. For example, Debarsy et al. (2012) generalize the spatial autoregressive model to incorporate time-lags in own area and neighboring areas, as in

$$y_{it} = \phi y_{i,t-1} + \lambda \sum_j w_{ij} y_{jt} + \theta \sum_j w_{ij} y_{j,t-1} + X_{it}\beta + \gamma \sum_j w_{ij} X_{jt} + \alpha + \varepsilon_{it}$$

This model does not include permanent area effects, whereas Kakamu and Wago (2008) propose

$$y_{it} = \lambda \sum_j w_{ij} y_{jt} + X_{it}\beta + \alpha_i + u_{it}$$

where u_{it} is iid, $u_{it} \sim N(0, \sigma^2)$. Assuming α_i is random, for example $\alpha_i \sim N(0, \sigma_\alpha^2)$, the stage 1 likelihood for period t is

$$p(y_t | \lambda, \beta, \alpha, \sigma^2) = (2\pi\sigma^2)^{-n/2} |I - \lambda W| \exp(-0.5u_t' u_t / \sigma^2)$$

where $u_t = y_t - \lambda W y_t - X_t \beta - \alpha$, and α is the vector of area permanent effects. As for cross-sectional spatial lag or error models, computational savings are achieved by taking uniform grid priors on λ , allowing pre-calculation of the log determinants, $\log|I - \lambda W|$, at each grid point.

Conditional hierarchical space-time priors may have benefits in MCMC applications and are applicable straightforwardly in area-time analysis involving binomial or Poisson count data. For example, in Poisson modelling of area health risks ρ_{it} , one may, by analogy to the random intercept-random slope model of conventional panel models, assume spatially structured area-specific random variation for both the level and the growth effect, so that neighboring areas have similar trends in relative risk (Bernardinelli et al. 1995). For equally spaced time points and expected events E_{it} , one has

$$\begin{aligned} y_{it} &\sim Po(E_{it}\rho_{it}) \\ \log(\rho_{it}) &= \alpha + \delta_t + \lambda_{1i} + \lambda_{2it} \end{aligned}$$

where the level effects λ_{1i} describe the stable relative risk pattern, while trend parameters λ_{2i} describe incremental changes in relative risk. The broad scale trend is represented by parameters δ_t , which for T small may be modelled as fixed effects

with a corner constraint (e.g., $\delta_1 = 0$). The two sets of spatial effects $\{\lambda_{1i}, \lambda_{2i}\}$ can be assigned a bivariate conditional autoregressive prior, $IMCAR(\rho, 2)$, as discussed above.

To allow for local heterogeneity, space-time priors can incorporate the convolution principle, combining a pure spatial signal with an *iid* term, as in

$$\log(\rho_{it}) = \alpha + \delta_t + \omega_{1i} + u_{1i} + (\omega_{2i} + u_{2i})t$$

where u_{1i} and u_{2i} are *iid*, and $(\omega_{1i}, \omega_{1i})$ are separate *ICAR*, or jointly *IMCAR* with $\rho = 1$. Setting $c_{ji} = \omega_{ji} + u_{ji}$ one has

$$\log(\rho_{it}) = \alpha + \delta_t + c_{1i} + c_{2i}t$$

While some realignment of spatial risks is likely over time, one may, however, seek to model persistent differentials. Let $c_{it} = \omega_{it} + u_{it}$ denote a convolution scheme, combining area-time specific spatial effects ω_{it} and *iid* effects, u_{it} . Then, correlation through time can be represented by an AR1 process, with

$$\log(v_{it}) = \alpha + \delta_t + c_{it} + \lambda c_{i,t-1}, \quad \lambda \in (-1, 1), \quad t > 1$$

with initial time model (at $t = 1$) being

$$\log(v_{i1}) = \alpha + \delta_1 + \frac{c_{i1}}{(1 - \lambda^2)^{0.5}}$$

Space and time dependence in area-time interactions c_{it} can also be represented using a Kronecker product of the relevant structure matrices defining the inverse covariance matrices in the joint prior (Lagazio et al. 2001). Thus, an *ICAR(1)* scheme for spatial errors, with interaction matrix C based on adjacency, has a joint multivariate normal prior with inverse covariance $\tau_s K_s$, where τ_s is a precision parameter and the off-diagonal terms $K_{s[ij]}$ are -1 for neighboring areas i and j , and $K_{s[ij]} = 0$ otherwise. Diagonal terms in K_s are given by M_i , the number of neighbors of area i . For time, one may assume a low-order random walk (RW) prior. If a first-order RW prior in time is assumed with K_t as the structure matrix in the joint prior, then the off-diagonal elements are $K_{t[ab]} = -1$ for adjacent times a and b , and $K_{t[ab]} = 0$ otherwise. Diagonal terms equal 1 when $a = b = 1$ or $a = b = T$, and equal 2 for other diagonal terms. Then, an area-time interaction effect c_{it} formed by crossing an *RW1* time prior with a *ICAR(1)* spatial effect has a joint prior with precision specified by the Kronecker product

$$\tau_c K_s \otimes K_t$$

The corresponding conditional priors (for c_{it} conditioning on all other interactions) have precisions $\tau_c M_i$ when $t = 1$ or $t = T$, and $2\tau_c M_i$ otherwise. With L_i denoting the neighborhood of area i , the prior conditional means \bar{c}_{it} for c_{it} are

$$\begin{aligned}\bar{c}_{i1} &= c_{i2} + \sum_{j \in L_i} \frac{c_{j1}}{M_i} - \sum_{j \in L_i} \frac{c_{j2}}{M_i} \\ \bar{c}_{it} &= 0.5(c_{i,t-1} + c_{i,t+1}) + \sum_{j \in L_i} \frac{c_{jt}}{M_i} - \sum_{j \in L_i} \frac{(c_{j,t+1} + c_{j,t-1})}{(2M_i)}, \quad 1 < t < T \\ \bar{c}_{iT} &= c_{i,T-1} + \sum_{j \in L_i} \frac{c_{jT}}{M_i} - \sum_{j \in L_i} \frac{c_{j,T-1}}{M_i}\end{aligned}$$

For identification, the c_{it} should be doubly centered at each iteration (over areas for a given t and over times for a given area i).

72.6 Focused Clustering Models

In environmental epidemiology, disease risk may be related to proximity to one or more known or unknown hazard sites (e.g., Ismaila et al. 2007; Maule et al. 2007). A benchmark scheme in such situations includes background risk and focused risk (Diggle 1990), with relative risk for subjects at location s in relation to a point source at s_0 represented as

$$\lambda(s, s_0) = \rho g_0(s)g_1(s, s_0)$$

where ρ is the regional incidence rate, $g_0(s)$ is the population at risk at location s (or more broadly the background risk), and $g_1(s, s_0)$ expresses disease exposure postulated to reflect location in relation to the source. For example, one may take $g_1(s, s_0)$ to be a function of distance $d = |s - s_0|$ from the source (so that direction has no impact), namely,

$$g_1(s, s_0) = g_1(d) = 1 + \eta f(d, \phi)$$

where $f(d, \phi)$ is a distance-decay function expressing lessened risk at greater distance, such as an exponential function, $f(d) = \exp(-\phi d)$, and where $1 + \eta$ defines relative risk at or near the source (where $d \simeq 0$ and $f(d) \simeq 1$). Provided $\phi > 0$, $g_1(d)$ tends to 1 as d tends to infinity (and $f(d)$ tends to zero). Probabilities of excess risk for particular subjects (at distance d_i from the source) may be obtained by monitoring

$$I(g_1^{(t)}(d_i) > g_E)$$

over MCMC iterations t , where g_E is judged to represent excess risk depending on the context (e.g., $g_E = 1.25$ or $g_E = 1.5$). The posterior probability estimate is then $\sum_{t=1}^T I(g_1^{(t)}(d_i) > g_E)/T$ where T is the total number of iterations. A simpler “hot spot”

clustering model specifies uniformly elevated risk $1 + \eta$ in a neighborhood (defined by distances $d < \delta$ around the focus, but background risk elsewhere. If there are multiple foci, one may generalize to

$$\lambda(s, s_0) = \rho g_0(s) \left[1 + \sum_k \eta_k f(d, \phi_k) \right]$$

Observed data for focused clustering models may involve individual-level disease status or small-area disease totals. For the former type of outcome, the population density may be modelled via kernel methods, for example, using small-area population estimates. An alternative is to proxy the background population distribution using a control disease unrelated to exposure from the point source. Cases and controls have binary outcomes $y_i = 1$ and $y_i = 0$, respectively, and if there are individual risk factors X_i , the odds of being a case may be represented as

$$\pi_i / (1 - \pi_i) = \rho^* [1 + \eta f(d)] \exp(X_i \beta)$$

where $\rho^* = (a/b)\rho$, a and b are sampling proportions of cases and of controls, respectively, and ρ is the population odds of disease.

Focused clustering may be relevant to small-area studies, for example, in modelling area counts of cancer incidence according to distance from a source or in modelling human flow behaviors (to hospitals or supermarkets). For observations consisting of disease counts y_i in areas i , the background risk in an area might be approximated by the expected disease total E_i based on population totals or age structure. To account for spatial correlation effects distinct from the effect of distance from the focus (or foci), the Poisson mean μ_i might include conditionally autoregressive spatial effects ϵ_i as discussed above, area predictors (e.g., deprivation), and *iid* effects u_i also, as in

$$\mu_i = \rho E_i \exp(X_i \beta) [1 + f(d_i)] \exp(u_i + \epsilon_i)$$

For example, setting $\alpha = \log(\rho)$, an exponential decay model would lead (Ma et al. 2007) to

$$\log(\mu_i) = \alpha + \log(E_i) + X_i \beta + \log(1 + e^{-\phi d}) + u_i + \epsilon_i$$

72.7 Conclusions

The chapter has reviewed some themes relevant to Bayesian applications and inferences to spatial data. There are many further issues to consider such as the development of efficient MCMC sampling for certain types of spatial model or in large datasets (e.g., Murray et al. 2010) and also the development of approximate Bayesian estimation methods (Rue et al. 2009). Among areas offering potential for

methodological development and benefitting from a Bayesian inferential perspective are space-time models used in econometric, health and climate applications (Tingley and Huybers 2010), nonparametric models for spatial data (e.g., Reich and Fuentes 2012), models for spatial cluster detection, and more general models for the spatial interaction matrix beyond simple assumptions such as contiguity. Routine application of Bayesian techniques to spatial and space-time models also depends on the availability of suitable software, and this is exemplified by packages developed for the freeware R package, such as Ramps and spBayes.

References

- Anselin L, Bera A (1998) Spatial dependence in linear regression models with an introduction to spatial econometrics. In: Ullah A, Giles D (eds) *Handbook of applied economic statistics*. Marcel Dekker, New York, pp 237–289
- Beck N, Gleditsch K, Beardsley K (2006) Space is more than geography: using spatial econometrics in the study of political economy. *Int Stud Quar* 50:27–44
- Bell B, Broemeling L (2000) A Bayesian analysis for spatial processes with application to disease mapping. *Stat Med* 19:957–974
- Bernardinelli L, Clayton D, Pascutto C, Montomoli C, Ghislandi M (1995) Bayesian analysis of space-time variations in disease risk. *Stat Med* 11:983–1007
- Besag J, Kooperberg C (1995) On conditional and intrinsic autoregressions. *Biometrika* 82(4):733–746
- Carroll C, Johnson D, Dunk J, Zielinski W (2010) Hierarchical Bayesian spatial models for multispecies conservation planning and monitoring. *Conserv Biol* 24:1538–1548
- Casella G, George E (1992) Explaining the Gibbs sampler. *Am Stat* 46:167–174
- Debarsy N, Ertur C, Sage J (2012) Interpreting dynamic space-time panel data models. *Stat Methodol* 9:158–171
- Diggle P (1990) A point process modelling approach to raised incidence of a rare phenomenon in the vicinity of a prespecified point. *J Roy Stat Soc* 153:349–362
- Diggle P, Tawn J, Moyeed R (1998) Model-based geostatistics. *J Roy Stat Soc C* 47:299–350
- Diggle P, Ribeiro P (2007) *Model-based Geostatistics*. Springer, New York
- Ecker M, Gelfand A (1997). Bayesian variogram modeling for an isotropic spatial process. *J Agric, Biol, Environ Stat*
- Goovaerts P, Gebreab S (2008) How does Poisson kriging compare to the popular BYM model for mapping disease risks? *Int J Health Geogr* 7:6
- Higdon D (2007) A primer on space-time modelling from a Bayesian perspective, Chapter 6. In: Finkelstadt BF, Held L, Isham V (eds) *Statistical methods for spatio-temporal systems*. CRC Press, Boca Raton
- Hossain M, Lawson A (2006) Cluster detection diagnostics for small area health data. *Stat Med* 25:771–786
- Irvine K, Gitelman A, Hoeting J (2007) Spatial design and properties of spatial correlation: effects on covariance estimation. *J Agric Biol Environ Stat* 12(4):1–20
- Ismaila A, Canty A, Thabane L (2007) Comparison of Bayesian and frequentist approaches in modelling risk of preterm birth near the Sydney Tar Ponds, Nova Scotia Canada. *BMC Med Res Methodol* 7:379
- Kakamu K, Wago H (2008) Small-sample properties of panel spatial autoregressive models: comparison of the Bayesian and maximum likelihood methods. *Spatial Econ Anal* 3:305–319
- Lagazio C, Dreassi E, Bernardinelli A (2001) A hierarchical Bayesian model for space-time variation of disease risk. *Stat Model* 1(17):29
- Lawson A, Clark A (2002) Spatial mixture relative risk models applied to disease mapping. *Stat Med* 21:359–370

- Lee D (2011) A comparison of conditional autoregressive models used in Bayesian disease mapping. *Spatial Spatio-Temporal Epidemiol* 2:79–89
- Lesage J, Pace K (2009) Introduction to spatial econometrics. CRC Press, Boca Raton
- Ma B, Lawson A, Liu Y (2007) Evaluation of Bayesian models for focused clustering in health data. *Environmetrics* 18:871–887
- Mardia K (1988) Multi-dimensional multivariate Gaussian Markov random fields with application to image processing. *J Multiv Anal* 24:265–284
- Maule M, Magnani C, Dalmasso P, Mirabelli D, Merletti F, Biggeri A (2007) Modeling mesothelioma risk associated with environmental asbestos exposure. *Environ Health Perspect* 115:1066–71
- Mollie A (1996) Bayesian mapping of disease. In Gilks W, Richardson S, Spiegelhalter DJ (eds) *Markov chain Monte Carlo in practice*. Chapman & Hall, London, pp 359–379
- Murray I, Prescott Adams R, MacKay D (2010) Elliptical slice sampling. *J Mach Learn Research - Proc Track* 2010:541–548
- Ratcliffe J (2010) Crime mapping: spatial and temporal challenges. In: Piquero A, Wiesburd D (eds) *Quantitative criminology*. Springer, New York, pp 5–24
- Reich B, Fuentes M (2012) Nonparametric Bayesian models for a spatial covariance. *Stat Methodol* 9:265–274
- Richardson S, Thomson A, Best N, Elliott P (2004) Interpreting posterior relative risk estimates in disease-mapping studies. *Environ Health Perspect* 112:1016–25
- Rodrigues A, Assuncao R (2008) Propriety of posterior in Bayesian space varying parameter models with normal data. *Stat Probab Lett* 78:2408–2411
- Rue H, Martino S, Chopin N (2009) Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *J Roy Stat Soc B* 71:319–392
- Schur N, Hürlimann E, Stensgaard A, Chimfwenbe K, Mushinge G, Simoonga C, Kabatereine N, Kristensen T, Utzinger J, Vounatsou P (2011) Spatially explicit Schistosoma infection risk in eastern Africa using Bayesian geostatistical modelling. *Acta Trop* 2011 Oct 14 (Epub ahead of print)
- Smith T, LeSage J (2004) A Bayesian probit model with spatial dependencies. In J LeSage R, Pace K (eds) *Advances in econometrics: volume 18: spatial and spatiotemporal econometrics*. Elsevier, Oxford, pp 127–160
- Stern H, Cressie N (2000) Posterior predictive model checks for disease mapping models. *Stat Med* 19:2377–2397
- Sun D, Tsutakawa R, Speckman P (1999) Posterior distribution of hierarchical models using CAR (1) distributions. *Biometrika* 86(2):341–350
- Wakefield J, Morris S (2001) The Bayesian modeling of disease risk in relation to a point source. *J Am Stat Assoc* 96:77–91
- Tingley M, Huybers P (2010) A Bayesian algorithm for reconstructing climate anomalies in space and time. Part I: development and applications to paleoclimate reconstruction problems. *J Climate* 23:2759–2781
- Wheeler D, Waller L (2009) Comparing spatially varying coefficient models: a case study examining violent crime rates and their relationships to alcohol outlets and illegal drug arrests. *J Geogr Syst* 11:1–22

David C. Wheeler

Contents

73.1	Introduction	1436
73.2	Model Specification	1436
73.3	Model Estimation	1439
73.4	Implementation	1442
73.5	Issues	1442
73.5.1	Statistical Inference	1442
73.5.2	Collinearity	1443
73.6	Diagnostic Tools	1443
73.7	Extensions	1446
73.8	Alternatives	1447
73.9	Application: Residential Chlordane Exposure in Los Angeles County	1449
73.10	Conclusions	1457
	References	1458

Abstract

Geographically weighted regression (GWR) was proposed in the geography literature to allow relationships in a regression model to vary over space. In contrast to traditional linear regression models, which have constant regression coefficients over space, regression coefficients are estimated locally at spatially referenced data points with GWR. The motivation for the introduction of GWR is the idea that a set of constant regression coefficients cannot adequately capture spatially varying relationships between covariates and an outcome variable. GWR is based on the appealing idea from locally weighted regression of estimating local models for curve fitting using subsets of observations centered on a focal point. GWR has been applied widely in diverse fields, such as ecology,

D.C. Wheeler

Department of Biostatistics, Virginia Commonwealth University, Richmond, VA, USA
e-mail: dcwheels@gmail.com; dcwheeler@vcu.edu

forestry, geography, and regional science. At the same time, published work from several researchers has identified methodological issues and concerns with GWR and has questioned the application of the method for inferential analysis. One of the concerns with GWR is with strong correlation in estimated coefficients for multivariate regression terms, which makes interpretation of map patterns for individual terms problematic. The evidence in the literature suggests that GWR is a relatively simple and effective tool for spatial interpolation of an outcome variable and a more problematic tool for inferring spatial processes in regression coefficients. The more complex approach of Bayesian spatially varying coefficient models has been demonstrated to better capture spatial nonstationarity in regression coefficients than GWR and is recommended as an alternative for inferential analysis.

73.1 Introduction

Geographically weighted regression (GWR) was proposed in the geography literature by Brunsdon et al. (1996) to allow relationships in a regression model to vary over space. In contrast to traditional linear regression models, where the regression coefficients are constant over space, regression coefficients are estimated locally at spatially referenced data points with GWR. The movement of local regression coefficients away from their global values, where the global values come from a traditional linear regression model, is termed parametric nonstationarity and spatial nonstationarity in the case of spatial processes (Brunsdon et al. 1996). The motivation for the introduction of GWR is the idea that it is unreasonable to assume that a set of constant regression coefficients can adequately capture spatially varying relationships between covariates and an outcome variable.

GWR is based on the simple idea of estimating local models using subsets of observations located around a focal point. GWR has as its methodological foundation the nonparametric technique of locally weighted regression, developed in statistics for curve-fitting and smoothing applications. In locally weighted regression, parameters are estimated using subsets of data proximate to a model estimation point in variable space, where observations in the subset are applied weights that decrease with increasing distance in variable space. The modification proposed with GWR is to use a subset of data proximate to the model estimation location in geographic space in place of variable space.

Though the emphasis with traditional locally weighted regression in statistics has been on curve fitting, i.e., predicting or estimating the outcome variable (Cleveland and Devlin 1988), GWR has been presented as a method for conducting statistical inference on spatially varying relationships, in an attempt to extend the original emphasis on prediction to confirmatory analysis. The use of GWR for inferential analysis has been questioned and criticized, however, and it has been suggested that the method is more appropriately used for interpolation of an outcome variable, which is more in harmony with its origins. This chapter reviews the details of specifying and estimating a geographically weighted regression

model, summarizes a few important concerns with GWR, presents some diagnostic tools and alternative approaches, and concludes with an illustrative analysis of estimating concentrations of the pesticide chlordane with GWR.

73.2 Model Specification

In notation, the foundation for GWR is the traditional linear regression model

$$y_i = \beta_0 + \sum_{k=1}^{p-1} \beta_k x_{ki} + \varepsilon_i \quad (73.1)$$

where y_i is the normally distributed outcome variable and x_{ki} is the value for the k th covariate for observation i , β_0 is the intercept, β_k is the regression coefficient for the k th covariate, and ε_i is the random error for observation i . There are p regression coefficients to estimate with the linear regression model. GWR specifies that the regression coefficients vary over observations as

$$y_i = \beta_{0i} + \sum_{k=1}^{p-1} \beta_{ki} x_{ki} + \varepsilon_i \quad (73.2)$$

where there is now an intercept and covariate regression coefficient for each data point. For $i = 1, \dots, n$ observations in the dataset, there are np regression coefficients to estimate because p coefficients are estimated at each of the n observations. It is required that the observations are spatially referenced, i.e., spatial coordinates are known to represent each data point. The observations may be areal units or individual-level data, such as residences. In the case of area data, the centroid of each areal unit is typically used as the basis for the spatial coordinates.

As with the linear regression model, it is convenient to express the GWR model in matrix notation

$$y_i = \mathbf{X}_i \boldsymbol{\beta}_i + \varepsilon_i \quad (73.3)$$

where \mathbf{X}_i is a row vector of explanatory variables and $\boldsymbol{\beta}_i$ is a column vector of regression coefficients at location i . In GWR, spatial structure is specified in the model through applying weights to the data. The weights are applied to the outcome variable and the covariates. The weights are calculated from a kernel function that typically assigns more weight to observations that are spatially closer to the data point (i th location) where the model is estimated. The introduction of the weights into the model follows from the assumption of spatial autocorrelation, where observations more proximate in space are thought to be more similar. Spatial autocorrelation ignored in the linear regression model results in spatially correlated errors, a violation of the model assumption of independent and identically

distributed errors. One can choose to model spatial autocorrelation either through the error term or through the regression coefficients, which is the GWR approach.

The kernel function used to calculate the weights in the GWR setting takes as input distances between all locations, conveniently in the form of a distance matrix. The kernel function has a bandwidth parameter that determines the spatial range of the kernel. The bandwidth parameter must be selected a priori or estimated from the data. The function returns a weight between locations that is inversely related to distance. A number of different kernel functions can be used in GWR. There are two general types of kernel functions, adaptive and fixed. Adaptive kernel functions inherently adjust for the density of data points by using a bandwidth expressed in number of observations. This results in a spatially larger kernel in data-sparse areas and a smaller kernel in data-dense areas. Fixed kernel functions have a bandwidth expressed as a constant distance; hence, the kernel is the same spatial size regardless of the density of data points. This could result in a varying number of observations weighted in kernels across the study area if the kernel function has weights that are zero beyond a certain distance.

Some of the most popular fixed kernel functions applied within GWR are continuous functions that produce weights that monotonically decrease with distance, such as the Gaussian or exponential kernel functions. The Gaussian kernel function is

$$w_{ij} = \exp\left(-\frac{1}{2}\left(\frac{d_{ij}}{\gamma}\right)^2\right) \quad (73.4)$$

where w_{ij} is the weight for data at location j in the model estimated for location i ; d_{ij} is the distance between locations i and j ; γ is the kernel bandwidth, a distance, that controls the decay and range of spatial correlation; and $\exp()$ is the exponential function. For matrix multiplications in the calculations of the model parameter estimates in GWR, the n weights for each model calibration location i , in row vector \mathbf{w}_i , are placed in an $n \times n$ weights matrix \mathbf{W} . The simpler exponential kernel function is

$$w_{ij} = \exp\left(-\frac{d_{ij}}{\gamma}\right) \quad (73.5)$$

which removes the scaling and powering of the Gaussian function. Another fixed kernel function is the bi-square kernel function

$$w_{ij} = \begin{cases} \left[1 - \left(d_{ij}^2/\gamma^2\right)\right]^2 & \text{if } d_{ij} \leq \gamma \\ 0 & \text{if } d_{ij} > \gamma \end{cases} \quad (73.6)$$

where the weight $w_{ij} = 0$ if the interpoint distance exceeds the kernel bandwidth γ . Hence, this function is continuous until a distance threshold is reached and then

is constant (zero) beyond the threshold. A similar kernel function is the tricube function

$$w_{ij} = \begin{cases} \left[1 - \left(d_{ij}^3/\gamma^3\right)\right]^3 & \text{if } d_{ij} \leq \gamma \\ 0 & \text{if } d_{ij} > \gamma \end{cases} \quad (73.7)$$

One of the more popular adaptive kernel functions is the bi-square nearest neighbor kernel. The function is

$$w_{ij} = \begin{cases} \left[1 - \left(d_{ij}/d_{iN}\right)^2\right]^2 & \text{if } j \text{ is one of the } N \text{th nearest neighbors of } i \\ 0 & \text{otherwise} \end{cases} \quad (73.8)$$

where d_{iN} is the distance to the N th nearest neighbor of location i and the number N of spatially nearest neighbors to use in the kernel function is estimated from the data. This function assigns a nonzero weight that decays with distance to points within the threshold number of neighbors and a weight of zero to points that are beyond the distance to the N th nearest neighbor.

Given the several options for a kernel function, one must first select the form of the kernel function before estimating the GWR model parameters, including the kernel bandwidth. Conventional thinking from the statistical nonparametric literature holds that the selection of the functional form for the kernel is less important than the selection of the kernel bandwidth for the model estimation results (Berk 2008). While this thinking is likely appropriate for GWR, a systematic assessment of the relative performance of different kernel functions in GWR has not been reported. Instead, most research has assumed a kernel function and focused on a criterion to select the kernel bandwidth.

73.3 Model Estimation

There are two methods for estimating the kernel bandwidth in GWR, cross-validation or minimizing the modified Akaike Information Criterion (AIC, Akaike 1973). Of the two approaches, cross-validation appears more commonly used, likely due to its heavy use in other related areas of statistics, such as local regression and statistical learning, and its conceptual simplicity.

Cross-validation is an iterative process that searches for the kernel bandwidth that minimizes the prediction error of all the observed outcome values using a subset of the data for prediction. The kernel bandwidth, denoted here generally as γ , is estimated in cross-validation by finding the γ that minimizes the cross-validation (CV) score. The sum of CV errors and the root mean squared prediction error (RMSPE) have been used as the CV score. The kernel bandwidth that minimizes the sum of CV errors, denoted $\hat{\gamma}$, is defined as

$$\hat{\gamma} = \arg \min_{\gamma} \sum_{i=1}^n [y_i - \hat{y}_{(i)}(\gamma)]^2 \quad (73.9)$$

where $\hat{y}_{(i)}$ is the predicted value of observation i with calibration location i left out of the estimation dataset. The kernel bandwidth minimizing the RMSPE is

$$\hat{\gamma} = \arg \min_{\gamma} \sqrt{\frac{1}{n} \sum_{i=1}^n [y_i - \hat{y}_{(i)}(\gamma)]^2} \quad (73.10)$$

This form of cross-validation is known as leave-one-out because only one observation is removed from the dataset for each local model when estimating the kernel bandwidth. The data point i is removed when estimating y_i to avoid estimating it perfectly. There are several search routines available for finding the optimal kernel bandwidth, including the golden search and the bisection search. Alternatively, one may systematically evaluate the CV score over a range of reasonable possible kernel bandwidths. In the kernel functions described above, the kernel bandwidth is a global parameter and is applied to all local models individually.

In contrast to cross-validation, the corrected AIC approach to estimating the kernel bandwidth is based on minimizing the estimation error of the outcome variable, not on the prediction of the outcome variable. The corrected AIC in GWR is adopted from locally weighted regression. The AIC is a compromise between model complexity and goodness of fit of the model, as there is a penalty for the effective number of parameters in the model. The corrected AIC for GWR is

$$AIC_c = 2n \log(\hat{\sigma}) + n \log(2\pi) + n \left(\frac{n + \text{trace}(\mathbf{H})}{n - 2 - \text{trace}(\mathbf{H})} \right) \quad (73.11)$$

where $\hat{\sigma}$ is the estimated standard deviation of the error, \mathbf{H} is the hat matrix, and the trace of a matrix is the sum of the matrix diagonal elements. The kernel bandwidth is used in the calculation of \mathbf{H} and $\hat{\sigma}$. Row i of the hat matrix is defined as

$$\mathbf{H}_i = \mathbf{X}_i (\mathbf{X}^T \mathbf{W}_i \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W}_i \quad (73.12)$$

which can also be expressed as

$$\mathbf{H}_i = \mathbf{X}_i \mathbf{A}_i \quad (73.13)$$

The estimated error variance is

$$\hat{\sigma}^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2 / (n - (2\text{trace}(\mathbf{H}) - \text{trace}(\mathbf{H}^T \mathbf{H}))) \quad (73.14)$$

To estimate the kernel bandwidth using the AIC, one can either use a search algorithm or evaluate the AIC over a range of possible bandwidth values to find the bandwidth that minimizes the AIC. The second approach is commonly used to show the relationship between the AIC and the kernel bandwidth.

The GWR estimates for the outcome variable values, \hat{y}_i , are calculated by

$$\hat{y}_i = \mathbf{X}_i \hat{\boldsymbol{\beta}}_i \quad (73.15)$$

where $\hat{\boldsymbol{\beta}}_i$ is a column of estimated regression coefficients at location i . The vector of estimated regression coefficients at one location is

$$\hat{\boldsymbol{\beta}}_i = [\mathbf{X}^T \mathbf{W}_i \mathbf{X}]^{-1} \mathbf{X}^T \mathbf{W}_i \mathbf{Y} \quad (73.16)$$

where $\mathbf{X} = [\mathbf{X}_1^T; \mathbf{X}_2^T; \dots; \mathbf{X}_n^T]^T$ is the design matrix of covariates and leading column of ones for the intercept, $\mathbf{W}_i = \text{diag}[w_{i1}, \dots, w_{in}]$ is the $n \times n$ diagonal weights matrix calculated for each location i , \mathbf{Y} is the $n \times 1$ vector of outcome variables, and $\hat{\boldsymbol{\beta}}_i = (\hat{\beta}_{0i}, \hat{\beta}_{1i}, \dots, \hat{\beta}_{pi})^T$ is the vector of p local regression coefficients at location i for $p - 1$ explanatory variables and the intercept. The weight matrix \mathbf{W}_i must be calculated at each location using the kernel function and bandwidth before the local regression coefficients can be estimated. The predictions of the outcome variable values $\hat{y}_{(i)}$ in cross-validation are calculated similarly, but with the element $w_{ii} \equiv 0$ to effectively remove the i th observation from consideration in the model to predict y_i . Given the definition of the estimated regression coefficients, GWR can be viewed as a locally weighted least squares regression model where the weights associate pairs of data points.

In previous studies using GWR, researchers have mapped estimated regression coefficients in attempts to interpret the spatial pattern of the coefficients in the context of the research problem (Brunsdon et al. 1996; Wheeler and Tiefelsdorf 2005). Researchers have typically been interested in where the estimated regression coefficients are statistically significant, according to some prespecified significance level. In the frequentist setting of traditional GWR, statistical significance tests of the coefficients use the variance of the estimated regression coefficients. According to Fotheringham et al. (2002, p. 55), the variance of the regression coefficients is

$$\text{Var}[\hat{\boldsymbol{\beta}}_i] = \mathbf{A}_i \mathbf{A}_i^T \hat{\sigma}^2 \quad (73.17)$$

Technically, this equation is incorrect because the Fotheringham et al. (2002) version of GWR is not a formal statistical model with kernel weights that are specified as part of the errors. The equation used for the local coefficient covariance is only approximate when using cross-validation because the kernel weights are calculated from the data first, before the regression coefficients are estimated from the data. The kernel weights are inherently a function of the outcome variable, as are the regression coefficients, and the correct expression for the coefficient covariance would be nonlinear.

73.4 Implementation

Implementations of GWR are freely available for the software R in the packages spgwr, written by Roger Bivand and Danlin Yu, and gwrr, written by the author. The package spgwr contains functions for estimating GWR model parameters by minimizing the AIC or using cross-validation with several possible kernel functions, including a Gaussian kernel, a bi-square kernel, and a tricube kernel. The gwrr package has functions for estimating GWR model parameters by cross-validation with exponential and Gaussian kernel functions and a bisection search routine. The gwrr package also has a function to diagnose collinearity with GWR models, and functions to estimate a penalized version of GWR, known as geographically weighted ridge regression, to dampen collinearity effects.

73.5 Issues

73.5.1 Statistical Inference

Though the introduction of GWR has provided an approach to investigate regression relationships that may vary over space, there are several critiques of the method. A central issue is a lack of formal statistical inference. GWR lacks a unified statistical framework, as it is effectively an ensemble of local spatial regressions, where the dependence between regression coefficients at different data locations is not specified in the model. This is a fixed effects model with no pooling across estimates. A consequence of a lack of a formal statistical model is that the standard error calculations in GWR are only approximate. This fact is due to reusing data for parameter estimation at multiple locations (Congdon 2003; Lesage 2004) and to using the data to estimate first the kernel bandwidth through cross-validation and then the regression coefficients (Wheeler and Calder 2007). The implication of the approximate standard errors is that the confidence intervals for estimated GWR coefficients are only approximate and should not be considered exactly reliable for detecting statistically significant covariate effects.

Another issue for inference on regression relationships with GWR is with the nature and amount of spatial variation in the estimated coefficients, i.e., nonstationarity. Tests for significant spatial variation in the estimated coefficients for one term in a GWR model have been proposed by Fotheringham et al. (2002) and Leung et al. (2000a). However, the tests do not consider the source of the spatial variation observed in the coefficients. There is concern that variation in the pattern of estimated coefficients may be artificially introduced by the smoothing methodology in GWR and may not represent true nonstationarity in the regression effects (Wheeler and Tiefelsdorf 2005). In other words, nonstationary regression effects could be an artifact of the methodology. Additionally, regression coefficient variability in GWR could result from collinearity effects.

In this light of uncertain statistical inference, GWR is more appropriately viewed as an exploratory approach and not a formal model to infer parameter nonstationarity. This view conflicts with the broad application of GWR as an inferential method. Instead of formal statistical inference on spatially varying regression effects, GWR is perhaps better suited to estimation and prediction of an outcome variable. This use would be more congruous with the theoretical origins of GWR in local linear regression, which was developed to estimate a response variable locally. GWR has produced favorable results in estimating a dependent variable compared with other interpolation techniques (Páez et al. 2008). Another argument for using GWR as a local estimator of a response variable is that when interpolation of an outcome variable over space is the main interest, regression coefficient estimation issues in GWR, such as collinearity, are no longer a major concern.

73.5.2 Collinearity

An issue that can interfere with statistical inference in linear regression models generally is collinearity. Collinearity is the presence of linear dependencies in the design matrix of a regression model, resulting in redundant information in the design matrix and an ill-conditioned variance matrix. Some of the negative consequences of collinearity are overestimates of covariate effect magnitudes, coefficient sign reversals, inflated variances for regression coefficients, and strong correlation in two or more estimated regression coefficients, all of which are likely to lead to incorrect interpretations of relationships in the regression model (Neter et al. 1996). These symptoms of collinearity have all been observed with GWR models for either simulated or actual datasets (Wheeler and Tiefelsdorf 2005; Waller et al. 2007; Griffith 2008; Finley 2011). The most conspicuous result of estimated GWR models pointing to collinearity effects in many studies has been strongly correlated regression coefficients for pairs of regression terms, including the intercept, evident in maps or scatter plots of estimated coefficients. Particular concern about collinearity symptoms is warranted with GWR, as collinearity has been found in empirical work to be an issue in local GWR models when it is not present in the traditional linear regression model with the same data (Wheeler 2007). Wheeler and Tiefelsdorf (2005) show through a simulation study that although GWR coefficients can be correlated when there is no correlation in explanatory variables, the coefficient correlation increases systematically with increasingly stronger collinearity.

73.6 Diagnostic Tools

There are several established diagnostic tools that have become an essential part of the practice of model fitting for traditional linear regression models, including methods to check for collinearity, influential observations, and autocorrelation. Use of a more complex regression model, such as GWR, should also be

complemented with diagnostic tools. Methods to identify spatial residual autocorrelation in GWR models have been developed by Leung et al. (2000b) and by Páez et al. (2002). A limitation of these approaches is that they are not model-based, remembering that the GWR method is a collection of local models that are not part of a unified framework. As a result, it is not clear that the source of autocorrelation can be identified. Farber and Páez (2007) proposed a method to adjust for influential observations in the cross-validation of GWR.

The need for diagnostic tools for collinearity in GWR models is motivated by several examples in the literature of the presence of redundant information in sets of estimated GWR coefficients from models built for different datasets (Wheeler and Tiefelsdorf 2005; Waller et al. 2007; Griffith 2008). Local collinearity in the GWR model can cause strong correlation in pairs of estimated regression coefficients, as is a consequence of collinearity in the traditional linear regression model. Based on existing published findings, researchers should strongly consider using diagnostic tools for collinearity when estimating GWR coefficients. Conveniently, there are diagnostic tools available to determine if there are substantial collinearity effects present in a GWR model. Simple tools to detect collinearity effects in GWR models include scatter plots of regression coefficients for pairs of regression terms, maps of approximate local regression coefficient correlations (Wheeler and Tiefelsdorf 2005), and local variance inflation factors (VIFs) (Wheeler 2007). The VIF measures how much the estimated variance of a regression coefficient is increased by collinearity. A limitation of the VIF as a diagnostic tool is that it does not consider collinearity with the intercept. More advanced and recommended tools are the variance-decomposition proportions and the associated condition indexes (Belsley 1991; Wheeler 2007). An advantage of the variance-decomposition approach over the VIF is that it measures and conveys the nature of the collinearity among all regression terms at the same time, including the intercept. As motivation for using these tools, applying them to a GWR model to explain crime rates in Columbus, OH, clearly linked local collinearity to strong GWR coefficient correlation and increased coefficient variation for two economic status covariates at numerous data locations with counterintuitive positive regression coefficient signs (Wheeler 2007).

The variance-decomposition proportion and condition index diagnostic tools introduced by Belsley (1991) and modified for GWR by Wheeler (2007) use singular value decomposition of the GWR kernel-weighted design matrix to calculate variance-decomposition proportions and condition indexes of the coefficient covariance matrix. The variance-decomposition proportion is the percentage of the variance of a regression coefficient that is explained with any one component of the variance matrix decomposition. The condition index is the ratio of the largest singular value and a smaller singular value of the decomposition. Each variance-decomposition proportion has an associated condition index. The singular value decomposition of the design matrix in the GWR model is

$$\mathbf{W}_i^{1/2} \mathbf{X} = \mathbf{UDV}^T \quad (73.18)$$

where \mathbf{U} and \mathbf{V} are orthogonal $n \times p$ and $p \times p$ matrices, respectively, and \mathbf{D} is a $p \times p$ diagonal matrix of singular values of $\mathbf{W}_i^{1/2}\mathbf{X}$, starting at matrix element (1,1) and decreasing in value down the diagonal. The matrix $\mathbf{W}_i^{1/2}$ is the square root of the diagonal weight matrix for calibration location i using a kernel function with the estimated kernel bandwidth from the GWR model. By way of the decomposition, the local variance-covariance matrix of the regression coefficients is

$$\text{Var}(\hat{\boldsymbol{\beta}}_i) = \sigma^2 \mathbf{V} \mathbf{D}^{-2} \mathbf{V}^T \quad (73.19)$$

The variance of the local k th regression coefficient is

$$\text{Var}(\hat{\beta}_{ik}) = \sigma^2 \sum_{j=1}^p \frac{v_{kj}^2}{e_j^2} \quad (73.20)$$

where the v_{kj} 's are the elements of the \mathbf{V} matrix and the e_j 's are the singular values. The variance-decomposition proportion for the local k th regression term and the j th component of the decomposition is

$$\pi_{jk} = \frac{\phi_{kj}}{\phi_k} \quad (73.21)$$

where

$$\phi_{kj} = \frac{v_{kj}^2}{e_j^2} \quad (73.22)$$

and

$$\phi_k = \sum_{j=1}^p \phi_{kj} \quad (73.23)$$

The condition index for variance component $j = 1, \dots, p$ is

$$\eta_j = \frac{e_{\max}}{e_j} \quad (73.24)$$

where e_{\max} is the largest singular value.

Belsley (1991) introduced guidelines for using the variance-decomposition proportions and condition indexes in the traditional linear regression setting. Through experimentation results, Belsley (1991) suggests a conservative value of 30 as a threshold for a condition index which indicates collinearity, although the threshold could be as low as 10 if there are large variance-decomposition proportions for two or more regression terms for the same variance component. In general, larger condition

indexes suggest stronger collinearity. A guideline for the variance-decomposition proportions is that the presence of two or more variance-decomposition proportions greater than 0.5 for the same variance component indicates that collinearity exists between those regression terms. It appears reasonable to apply these guidelines for diagnosing collinearity in a GWR model; however, the guidelines have not been systematically studied in the GWR setting.

The condition index and variance-decomposition proportion diagnostic tools reveal collinearity locally for the individual GWR models and consequently enable researchers to construct plots of the diagnostic values and link them directly to estimated GWR coefficients for visual analysis of any collinearity problems present in the model. Estimated GWR coefficients from local models that are diagnosed as problematic should be interpreted with severe caution, and additional analysis should be carried out in these areas to better understand the nature of the relationships being modeled. The variance-decomposition proportions and condition index tools are implemented in the freely available R package *gwrr*.

73.7 Extensions

Several different models have been introduced to extend the concept of geographically weighted regression. The focus here is on extensions that address the issue of collinearity. Collinearity effects in linear regression models have been damped by constraining the amount of variation in regression coefficients. Two extended versions of GWR have been proposed that are based on the coefficient shrinkage models of ridge regression and the lasso: geographically weighted ridge regression (GWRR; Wheeler 2007) and the geographically weighted lasso (GWL; Wheeler 2009). The methods limit the amount of variation in the coefficients through the addition of a constraint, or penalty, on the size of the regression coefficients. Ridge regression coefficients minimize the sum of a penalty on the residual sum of squares and the size of the squared coefficients,

$$\hat{\beta}^R = \arg \min_{\beta} \left\{ \sum_{i=1}^n \left(y_i - \beta_0 - \sum_{k=1}^p x_{ki} \beta_k \right)^2 + \lambda \sum_{k=1}^p \beta_k^2 \right\} \quad (73.25)$$

where λ is the shrinkage parameter controlling the magnitude of the regression coefficients. The lasso coefficients minimize the sum of the residual sum of squares and the absolute value of the coefficients,

$$\hat{\beta}^L = \arg \min_{\beta} \left\{ \sum_{i=1}^n \left(y_i - \beta_0 - \sum_{k=1}^p x_{ki} \beta_k \right)^2 + \lambda \sum_{k=1}^p |\beta_k| \right\} \quad (73.26)$$

The nature of the constraint in the lasso results in more shrinkage in the regression coefficients than with ridge regression, potentially shrinking some of

the coefficients to zero. The lasso solution coefficients are a transformation of the least squares solution by a constant factor, truncated at zero. As such, the lasso may be viewed as a model-selection method, in that it can remove a term from the model by zeroing its effect. Because the methods are scale dependent, it is common practice when using the lasso or ridge regression to center the response variable and center and scale the explanatory variables to have unit variances. The estimates of the GWRR coefficients using centering of the variables are

$$\hat{\beta}_i = (\mathbf{X}^{*T} \mathbf{W}_i \mathbf{X}^* + \lambda \mathbf{I})^{-1} \mathbf{X}^{*T} \mathbf{W}_i \mathbf{y}^* \quad (73.27)$$

where \mathbf{X}^* is the matrix of standardized explanatory variables, \mathbf{I} is the identity matrix, and \mathbf{y}^* is the standardized response variable. There are different options for the type of scaling and centering (see Wheeler 2007). The absolute value constraint on the regression coefficients in GWL makes the problem nonlinear, but computationally efficient algorithms exist for estimating the parameters (Wheeler 2009). GWRR is implemented in the R package gwrr.

73.8 Alternatives

Bayesian hierarchical models are an alternative to GWR for estimating spatially varying coefficients in regression models. The Bayesian hierarchical modeling framework is flexible and provides formal statistical inference for model parameters in a unified statistical model. In the Bayesian setting, the distribution of the outcome variable is specified conditional on unknown parameters, whose distribution is in turn conditional on other parameters. All parameters are considered random in a Bayesian model. As an alternative to GWR, one can use a Bayesian hierarchical model with random effects for the intercept and covariate effects and specify the random effects as either independent in the prior and borrow strength across observations globally or as spatially correlated and borrow strength locally. The second specification is closer in spirit to GWR, and Bayesian models of this type are called spatially varying coefficient (SVC) models. There are two different approaches for specifying the spatial structure in spatially varying coefficients (Banerjee et al. 2004). One is a prior conditional specification of the coefficients that uses only neighboring observations, and the other is a prior joint specification of the coefficients that models correlation in the coefficients as a continuous spatial process. The second specification is more similar in form to the fixed kernel functions described earlier for modeling spatial correlation in GWR.

The Bayesian SVC model with the continuous spatial process prior for the regression coefficients may be specified conveniently with matrix notation as

$$[\mathbf{Y} | \boldsymbol{\beta}_P, \tau^2] \tilde{N}(\mathbf{X}_P^T \boldsymbol{\beta}_P, \tau^2 \mathbf{I}) \quad (73.28)$$

where \mathbf{Y} is assumed to be Gaussian conditional on the parameters τ^2 and $\boldsymbol{\beta}_P$. $\boldsymbol{\beta}_P$ is an $np \times 1$ vector of regression coefficient parameters, and \mathbf{X}_P^T is the $n \times np$ block

diagonal matrix of covariates, where each row contains a row from the $n \times p$ design matrix \mathbf{X} along with zeros in the appropriate places (the covariates from \mathbf{X} are shifted p places in each subsequent row in \mathbf{X}_P^T). The subscript P denotes the different sizes of the design matrix and the regression coefficient matrix associated with the process model. τ^2 is the error variance and \mathbf{I} is the $n \times n$ identity matrix.

The prior distribution for the regression coefficient parameters is

$$[\boldsymbol{\beta}_P | \boldsymbol{\mu}_{\boldsymbol{\beta}}, \Sigma_{\boldsymbol{\beta}}] = N(\mathbf{1}_{n \times 1} \otimes \boldsymbol{\mu}_{\boldsymbol{\beta}}, \Sigma_{\boldsymbol{\beta}}) \quad (73.29)$$

where the vector $\boldsymbol{\mu}_{\boldsymbol{\beta}} = (\mu_{\beta_0}, \dots, \mu_{\beta_p})^T$ contains the means of the regression terms. The Kronecker product operator (\otimes) multiplies every element in $\mathbf{1}_{n \times 1}$ by $\boldsymbol{\mu}_{\boldsymbol{\beta}}$. The prior on the regression coefficients takes into account possible spatial dependence in the coefficients through the covariance, $\Sigma_{\boldsymbol{\beta}}$, which has a separable form with two components, one for the within-site dependence between coefficients and one for the spatial dependence in the regression coefficients. The separable form of the covariance matrix for $\boldsymbol{\beta}_P$ is

$$\Sigma_{\boldsymbol{\beta}} = \mathbf{R}(\phi) \otimes \mathbf{T} \quad (73.30)$$

where $\mathbf{R}(\phi)$ is the $n \times n$ correlation matrix that models the spatial association between the n locations using interpoint distances, ϕ is an unknown spatial dependence parameter, and \mathbf{T} is a positive-definite $p \times p$ matrix for the covariance of the regression coefficients at any spatial location. In contrast to the repeated application of spatial kernel functions in GWR, this $np \times np$ covariance matrix simultaneously captures the covariation between all the regression coefficients. In the separable covariance matrix, each of the p coefficients represented in the covariance is assumed to have the same spatial dependence structure. This matches the assumption in GWR of equivalent spatial ranges for each regression term.

The specification of the Bayesian SVC process model is completed with the specification of prior distributions for the other parameters. Conjugate priors are Gaussian for the coefficient means, inverse gamma for the error variance, and inverse Wishart for the within-site covariance matrix. The spatial dependence parameter can be specified with a uniform or gamma prior distribution. Inference on the model parameters is realized through Markov chain Monte Carlo (MCMC) by sampling from the joint posterior distribution of the parameters and then summarizing the distribution with posterior means, medians, and credible intervals. Details for the MCMC simulation are provided in Wheeler and Calder (2007) and Finley (2011). A variety of Bayesian spatial process models are implemented in the R package spBayes, written by Andrew Finley, Sudipto Banerjee, and Brad Carlin. Bayesian SVC models with spatial dependence specified through neighborhood adjacency may be fitted with GeoBUGS (Thomas et al. 2004) and WinBUGS software (Spiegelhalter et al. 2003).

Researchers can choose to model spatial autocorrelation in a regression model through the regression coefficients, as in GWR and the Bayesian SVC model, or through the error term. As such, assuming nonstationarity in the regression

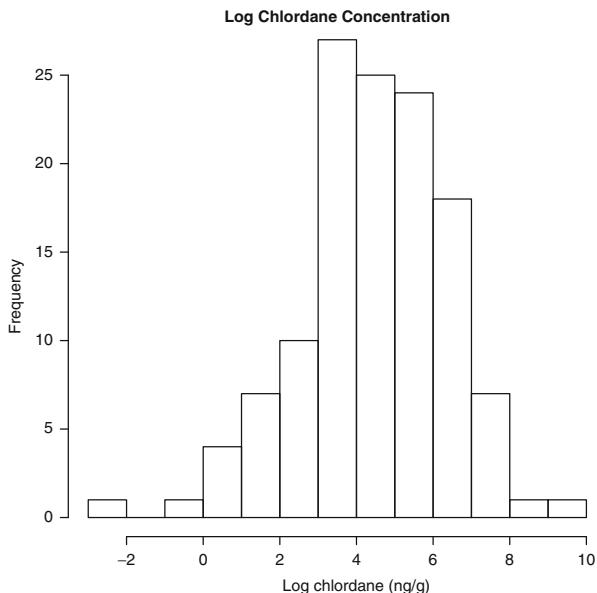
coefficients is a choice made about the source and form of the spatial dependence in the data. GWR was designed to address spatial autocorrelation through nonstationary regression coefficients. If one is not confident about the presence of spatially varying effects but wants to adjust for spatial autocorrelation, one can always specify a regression model with spatially structured residuals (see Congdon 2010 for examples in the Bayesian framework).

73.9 Application: Residential Chlordane Exposure in Los Angeles County

Insecticides are biologically active chemicals that have been used for agricultural and household applications for decades in the United States. Epidemiologic studies have linked residential exposure to insecticides with elevated risk of cancer, including non-Hodgkin lymphoma (NHL) (Colt et al. 2005, 2006). Specifically, positive associations were found between NHL risk and dust residues of dichlorodiphenyldichloroethylene (DDE), a metabolite of DDT (dichlorodiphenyltrichloroethylene) (Colt et al. 2005) and chlordane, which was used in the United States for subterranean termite control from 1978 to 1988 (Colt et al. 2006). Chlordane was banned in 1988 because of concerns over cancer risk and danger to wildlife. Risk of NHL was elevated among subjects who had home termite treatments before 1988, but not after 1988 (Colt et al. 2006). As with other organochlorine chemicals such as polychlorinated biphenyl (PCB), chlordane is persistent in the environment, and it has been detected in the indoor air of homes 15 years after termite treatment (Livingston and Jones 1981). Given the health concerns and the persistence of these organochlorine chemicals, it is important to describe their spatial distribution in the environment and determine factors that explain the distribution.

The National Cancer Institute Surveillance, Epidemiology, and End Results (NCI-SEER) NHL study was a multicenter, population-based case–control study that was conducted to investigate potential risk factors for NHL. Details of the study design have been reported previously (Colt et al. 2005; Wheeler et al. 2011). NHL cases 20–74 years of age were identified between July 1, 1998, and June 30, 2000, from the SEER registries covering Iowa and the metropolitan areas of Detroit, Seattle, and Los Angeles County. Controls were selected by random digit dialing or from Medicare files. Computer-assisted personal interviews (1,321 cases and 1,057 controls) were conducted in 1998–2000 to elicit data on medical history, demographic variables, and various risk factors for NHL, including home insecticide use. Participants were also asked to provide lifetime residential histories that were address matched to geographic address databases to yield spatial coordinates. For study subjects who had used a vacuum in the home in the past year and owned most of their carpets for 5 years or more (682 cases and 513 controls), carpet dust samples from used vacuum cleaner bags were collected to measure residential exposure to certain chemicals, including the insecticide chlordane. Some insecticides applied indoors or tracked in from the outdoors persist in carpet dust for months or years, where they are protected from degradation by rain, sunlight,

Fig. 73.1 Log chlordane concentration in Los Angeles County



temperature extremes, and microbial action. Pesticides and other chemicals in dust were analyzed using gas chromatography/mass spectrometry (ng per gram of fine dust [$<100\text{ }\mu\text{m}$ diameter]).

For this analysis, I used linear regression and geographically weighted regression models to explain chlordane levels in house dust in the NCI-SEER NHL study Los Angeles center, where chlordane levels were highest in the study. This example analysis could be read as a cautionary example in using GWR models; however, no data analysis or manipulation was performed to create an unfavorable situation in which to apply these models. Data on gender, race, education, home age, presence of Oriental rugs, housing type, and pest treatments of sampled homes were available from interviews. US Census 2000 tract variables for median income, median year homes were built, and percent of the tract that was urban were used to characterize the neighborhood socioeconomic status and urbanicity. The 1992 National Land Cover Dataset (NLCD) was used to determine land use around residences. Percent of land cover within specified distance buffers (500, 1,000, 2,000, 3,000, 5,000, 10,000, 15,000, 20,000 m) of each residence was calculated for classifications of use (agricultural, residential, urban/recreational, orchards/vineyards). All statistical analyses were performed using the software R.

Observed chlordane concentration was approximately lognormally distributed (Fig. 73.1); therefore, it was modeled as natural lognormal in the models. Chlordane concentrations below the minimum detection limit of 20.8 ng/g were imputed with a regression model (Colt et al. 2005). Concentrations of the insecticides α -chlordane and γ -chlordane were summed to compose the total chlordane

Table 73.1 Estimated parameters for the first traditional linear regression model and geographically weighted regression model for chlordane concentration in Los Angeles County

Coefficient	LM estimate	Standard error	p value	VIF	Mean GWR estimate
Intercept	119.40	25.93	<0.001		116.48
Race	0.90	0.38	0.019	1.21	0.85
Year built	-0.06	0.01	<0.001	1.07	-0.06
Termite treatment	1.10	0.30	<0.001	1.01	1.09
Median income	-4.3E-06	4.9E-06	0.378	1.19	-4.2E-06
F-statistic	10.02		<0.001		
R-squared	0.25				0.28 ^a
RMSE	1.64				1.60

^aR-squared is only approximate for GWR

VIF variance inflation factor, RMSE root mean squared error, LM linear regression model, GWR geographically weighted regression

concentration. The first step in the analysis was to use a traditional linear regression model of the form

$$y_i = \mathbf{X}_i \boldsymbol{\beta} + \varepsilon_i \quad (73.31)$$

where y_i is log chlordane concentration (ng/g) in the home at time of the study interview for subject i , \mathbf{X}_i is a row vector of explanatory variables for subject i , $\boldsymbol{\beta}$ is a column vector of regression coefficients, and ε_i is the random error for subject i . The first element of X_i is 1 for the intercept. The initial model considered the relationship of log chlordane concentration with the variables gender, race (whites and Asians vs. other races), education (<12 years, 12–15 years, ≥16 years), median year home built (tract level), median income (tract level), percent urban, presence of Oriental rug, home type, home termite treatment before 1988, lawn treatment, percent land used for agriculture within 15 km of residence, percent land used for orchards within 15 km, and percent residential land within 10 km. Lawn treatment was included as another potential source of organochlorine compounds. Of the covariates, only race, home termite treatment, and median year home built were statistically significant in the model at the 0.05 level. All other variables except median income were excluded from the model to reach a parsimonious model. Median income was retained in the model because it was suspected a priori to potentially have a significant relationship with chlordane concentration that could also vary over space. Lower income communities and poorly maintained housing are positively associated with pest infestation and consequently treatments for pests.

The estimated linear regression model parameters are listed in Table 73.1. Median year home built and home termite treatment were highly statistically significant with intuitive coefficient signs: newer homes were associated with lower concentrations of chlordane, and termite treatments before 1988 was associated with increased chlordane concentrations. Whites and Asians had lower

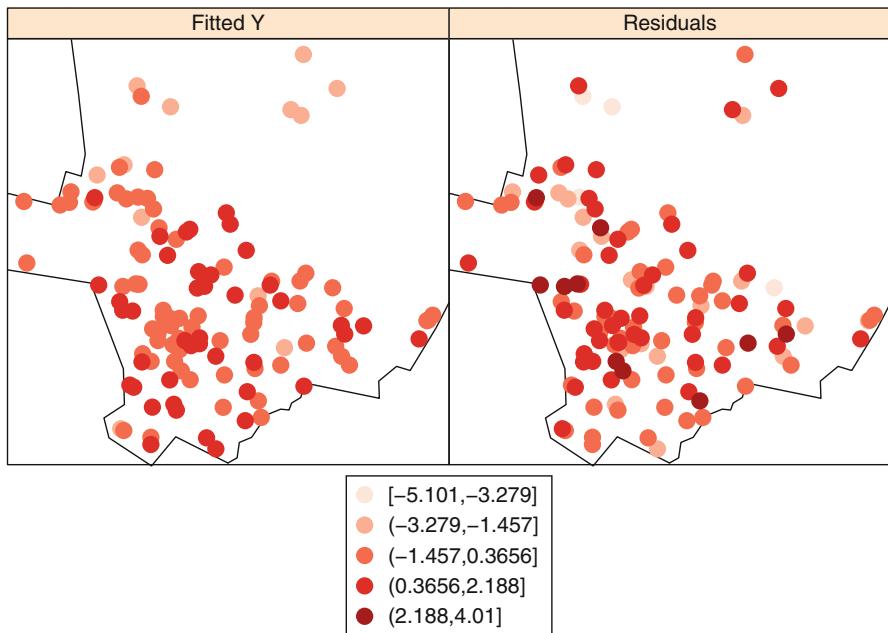


Fig. 73.2 Fitted log chlordane concentration and model residuals from the traditional linear regression model

residential chlordane concentrations than other races. Increasing median income was associated with lowered chlordane concentration, although not significantly. The coefficient of determination (R^2) was 0.25, and the root mean squared error (RMSE) was 1.64. The variance inflation factors did not suggest a problem of collinearity for any covariate term with the other covariate terms. The standard error for the intercept appears large, and generally this could suggest a collinearity issue; however, the standard error is not large relative to the intercept. The residuals of the model do not suggest lack of normality or heteroscedasticity (not shown). There does appear to be some local spatial pattern in the model residuals (Fig. 73.2). A Moran's I test using inverse distance weighting and 75 % of the data closest to each point yielded a p value of 0.01, suggesting that the null hypothesis of no spatial autocorrelation can be rejected.

Given the apparent spatial correlation in the residuals, I next used a generalized additive model (GAM) (Wood 2006) to model the residual spatial variation in chlordane concentration. The GAM specification is

$$y_i = \mathbf{X}_i \boldsymbol{\beta} + z(s) \quad (73.32)$$

where $z(s)$ is a smooth function of the spatial coordinates at location s , and the other terms are as previously defined. The smoothing function models spatial pattern in

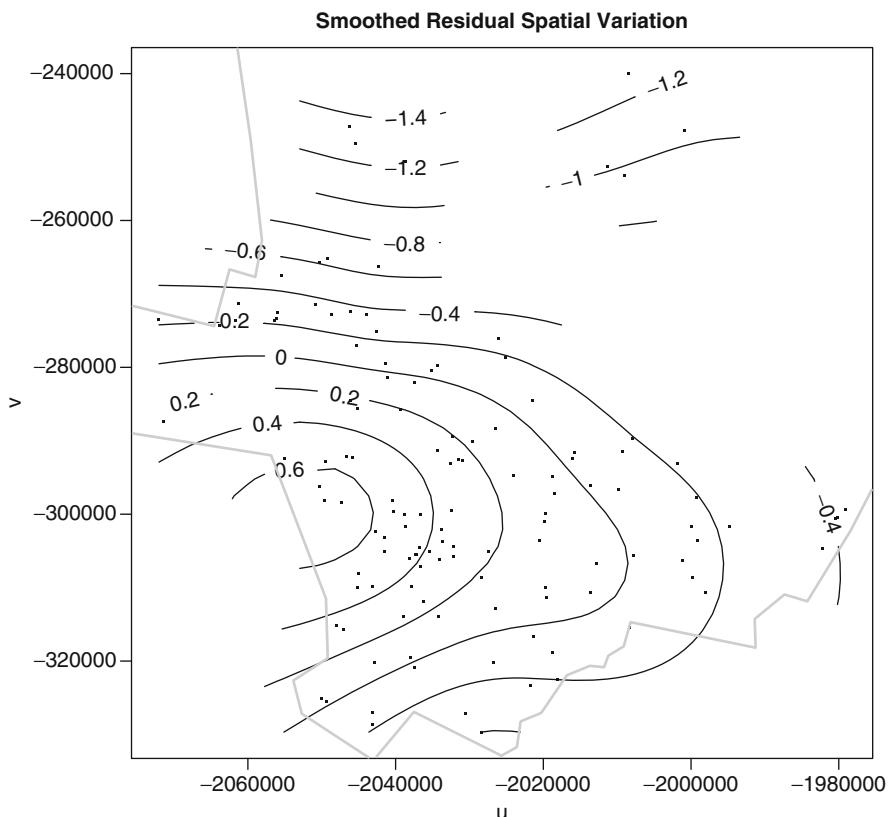


Fig. 73.3 Smoothed residual spatial variation from a generalized additive model for log chlordane concentration

what is not described by the covariates in the model. I used thin plate regression splines for the smoothing function (Wood 2006) with the R package mgcv. There is a clear pattern in the estimated residual spatial variation (Fig. 73.3), where residual chlordane concentration increases generally from southwest to northeast across the county. It is unknown whether the pattern observed in the residuals is due to spatial nonstationarity in the regression model relationships or from a misspecified model, where important covariates have been omitted.

To explore nonstationarity processes, I expanded on the basic linear regression model by estimating the GWR model specified in Eq. (73.3) for chlordane concentration using the R package gwrr. Though the GWR model improves slightly on the fit of the linear regression model (Table 73.1), the GWR coefficients are immediately suspicious. The scatter plot matrix of estimated regression coefficients shows that the coefficients for the intercept (β_0) and year home built (β_2) are nearly

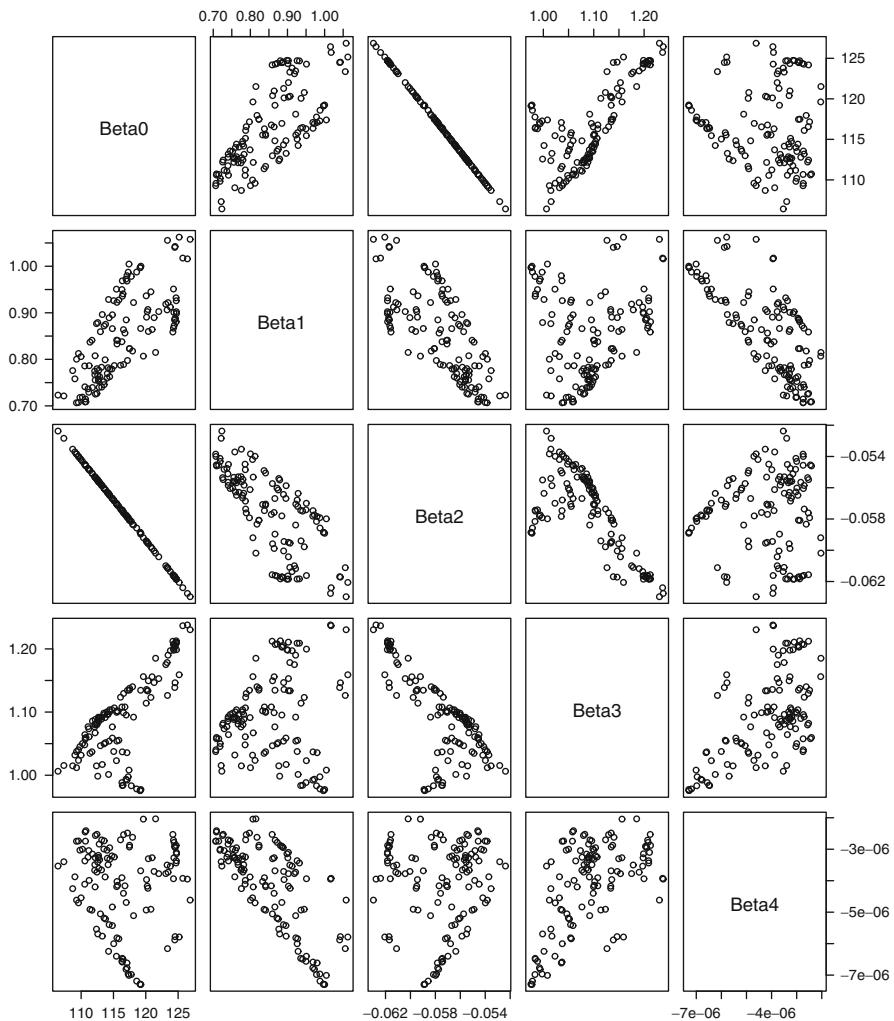


Fig. 73.4 Scatter plot matrix of estimated coefficients from the first GWR model

perfectly correlated (Fig. 73.4). The strong correlation in this pair of regression terms is evident in maps of the estimated coefficients (Fig. 73.5), where the maps are clearly complementary; β_0 is high where β_2 is low, and vice versa. The collinearity diagnostics for the GWR model indicate a major problem of collinearity with the intercept and year home built term. The condition index and variance-decomposition proportions calculated using the R package gwrr are extremely large, with the largest condition index for every local model exceeding 400 and the variance-decomposition proportions for the intercept and year home built each exceeding 0.99 for all local models. The strong collinearity results in

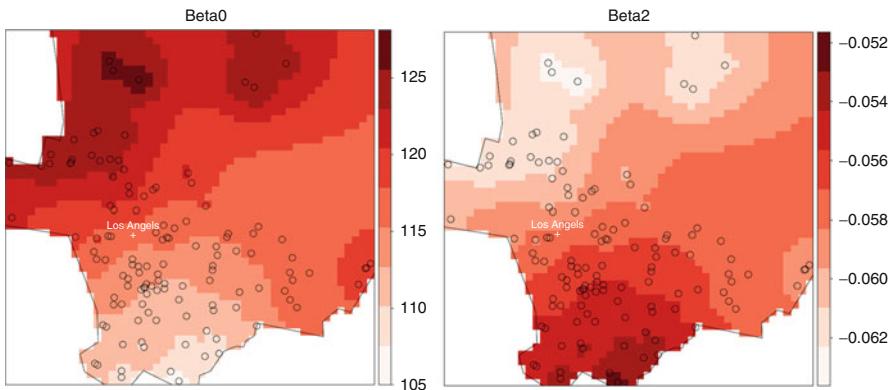


Fig. 73.5 Interpolated estimated regression coefficients for the intercept (Beta0) and year home built (Beta2) from the first GWR model with points indicating the sample locations

coefficients for year home built that are not meaningful, as they are perfectly correlated with the intercept, and knowing the local intercept effectively determines the year home built coefficient. Instead of having five regression terms, there are actually only four terms.

Reflecting back on the linear model, it would not be unexpected for practitioners to fail to detect an issue of collinearity in the model between the intercept and year home built, as some level of correlation with the intercept is expected with the linear model, researchers are not often encouraged to check for collinearity with the intercept, and the commonly used diagnostic of the variance inflation factor does not address collinearity with the intercept. Applying the collinearity diagnostics of the condition index and variance-decomposition proportions to the design matrix for the linear regression model revealed the collinearity issue, as the largest condition index was 504 and the corresponding variance-decomposition proportions for the intercept and year home built exceeded 0.99. Collinearity was present in the linear model, but it only became glaringly obvious with the GWR model. It would only have been detected using the appropriate, but infrequently applied, tools of the variance-decomposition proportions and condition index.

Potential solutions to the collinearity issue are to attempt a penalized regression or to remove the year home built variable from the model. I removed the year home built variable for both the linear regression model and the GWR model in this example. Removing this variable overcomes the collinearity issue in both models according to the collinearity diagnostics, as the condition indexes were on the order of 5 and there were no pairs of variance-decomposition proportions exceeding 0.5 for different regression terms. The fit of this second GWR model improved slightly by removing year home built, but the linear regression model fit decreased substantially (Table 73.2). In this case, when avoiding serious collinearity, GWR provided a substantially better fit than the linear regression model. Fitted values for chlordane concentration from the reduced GWR model convey the overall

Table 73.2 Estimated parameters for the reduced traditional linear regression model and geographically weighted regression model for chlordane concentration in Los Angeles County

Coefficient	LM estimate	Standard error	p value	VIF	Mean GWR estimate
Intercept	3.80	0.40	<0.001		3.92
Race	0.63	0.40	0.117	1.18	0.63
Termite treatment	1.23	0.32	<0.001	1.00	1.09
Median income	-6.4E-06	5.2E-06	0.218	1.18	-5.0E-06
F-statistic	5.82		<0.001		
R-squared	0.13				0.29 ^a
RMSE	1.77				1.59

^aR-squared is only approximate for GWR

VIF variance inflation factor, RMSE root mean squared error, LM linear regression model, GWR geographically weighted regression

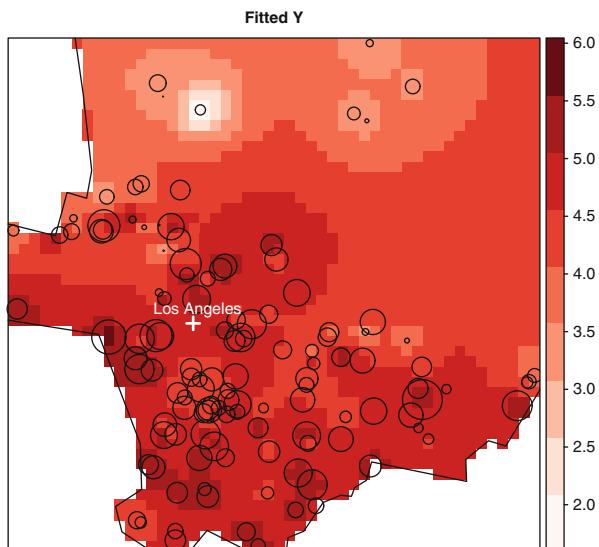


Fig. 73.6 Fitted chlordane concentration from the reduced GWR model (scale at right) with circles proportional to observed chlordane concentrations

pattern in the observed values, but are less granular (Fig. 73.6). There is noticeable spatial heterogeneity in the observed values, and the smoothing approach of GWR does not capture some of the local heterogeneity. Overall, the fitted and observed values are larger south of Los Angeles, although there is a line of higher concentration stretching northwest from Los Angeles into the San Fernando Valley. The estimated regression coefficients for the reduced GWR model still appear somewhat complimentary for pairs of regression terms, specifically for the intercept and home termite treatment, and for race and median income (Fig. 73.7). Even though the diagnostic tools do not indicate a problem with serious collinearity in the model, caution should still be used when interpreting the estimated regression coefficients.

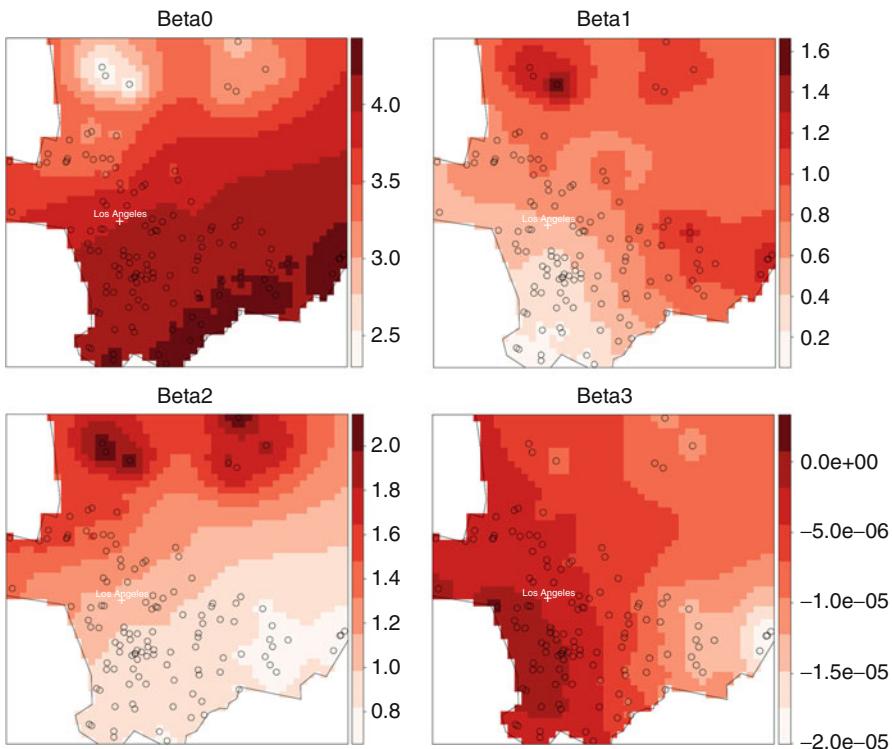


Fig. 73.7 Interpolated estimated regression coefficients from the reduced geographically weighted regression model. The points indicate the locations of sampled data. The regression coefficients are for the terms intercept (Beta0), race (Beta1), home termite treatment (Beta2), and median income (Beta3)

73.10 Conclusions

The main proposed application of geographically weighted regression for the spatial sciences was to investigate spatially varying relationships between variables in a regression model. GWR has been presented as a statistical model to be used for inference on spatially varying relationships. Based on a review of the literature, it has been adopted as such by researchers in several disciplines, including geography, regional science, ecology, and forestry. There is sufficient published evidence, however, to seriously question the uncritical application of GWR for studying multivariate spatially varying relationships. Evidence suggests that substantial spatial variation in estimated regression coefficients and correlation between estimated coefficients for pairs of regression terms is inherent to the method. Given the evidence, researchers should be extremely cautious about interpreting GWR coefficients and especially on making policy decisions based on GWR output. Instead, researchers should use diagnostic tools when building GWR models to detect

suspicious behavior of the model, including collinearity effects. When diagnostics reveal problems with the stability of the GWR model, alternative methods for making inferences about spatially varying processes should be considered. Alternatives include penalized versions of GWR and Bayesian spatially varying coefficient models, where Bayesian models provide formal statistical inference. Though the use of GWR for inference on spatially varying relationships has been questioned, it can be considered an accessible exploratory tool that is effective for estimating and predicting a spatially referenced outcome variable.

References

- Akaike H (1973) Information theory and an extension of the maximum likelihood principle. In: Petran B, Csaaki F (eds) International symposium on information theory. Akadeemiai Kiadó, Budapest, pp 267–281
- Banerjee S, Carlin BP, Gelfand AE (2004) Hierarchical modeling and analysis for spatial data. Chapman & Hall/CRC, Boca Raton
- Belsley DA (1991) Conditioning diagnostics: collinearity and weak data in regression. Wiley, New York
- Berk RA (2008) Statistical learning from a regression perspective. Springer, New York
- Brunsdon C, Fotheringham AS, Charlton ME (1996) Geographically weighted regression: a method for exploring spatial nonstationarity. *Geograph Anal* 28(4):281–298
- Cleveland WS, Devlin SJ (1988) Locally weighted regression: an approach to regression analysis by local fitting. *J Am Stat Assoc* 83:596–610
- Colt JS, Severson RK, Lubin L, Rothman N, Camann D, Davis S, Cerhan JR, Cozen W, Hartge P (2005) Organochlorines in carpet dust and non-Hodgkin lymphoma. *Epidemiology* 16(4):516–525
- Colt JS, Davis S, Severson RK, Lynch CF, Cozen W, Camann D, Engels EA, Blair A, Hartge P (2006) Residential insecticide use and risk of non-Hodgkin's lymphoma. *Cancer Epidemiol Biomark Prev* 15(2):251–257
- Congdon PD (2003) Modelling spatially varying impacts of socioeconomic predictors on mortality outcomes. *J Geograph Syst* 5(2):161–184
- Congdon PD (2010) Applied Bayesian hierarchical methods. Chapman & Hall/CRC, Boca Raton
- Farber S, Páez A (2007) A systematic investigation of cross-validation in GWR model estimation: empirical analysis and Monte Carlo simulations. *J Geograph Syst* 9(4):371–396
- Finley AO (2011) Comparing spatially-varying coefficients models for analysis of ecological data with non-stationary and anisotropic residual dependence. *Methods Ecol Evol* 2:143–154
- Fotheringham AS, Brunsdon C, Charlton M (2002) Geographically weighted regression: the analysis of spatially varying relationships. Wiley, West Sussex
- Griffith D (2008) Spatial filtering-based contributions to a critique of geographically weighted regression (GWR). *Environ Plan A* 40:2751–2769
- LeSage JP (2004) A family of geographically weighted regression models. In: Anselin L, Florax RJGM, Rey SJ (eds) Advances in spatial econometrics. Methodology, tools and applications. Springer, Berlin, pp 241–264
- Leung Y, Mei CL, Zhang WX (2000a) Statistical tests for spatial nonstationarity based on the geographically weighted regression model. *Environ Plan A* 32(1):9–32
- Leung Y, Mei CL, Zhang WX (2000b) Testing for spatial autocorrelation among the residuals of the geographically weighted regression. *Environ Plan A* 32(5):871–890
- Livingston JM, Jones CR (1981) Living area contamination by chlordane used for termite treatment. *Bull Environ Contam Toxicol* 27:406–411

- Neter J, Kutner MH, Nachtsheim CJ, Wasserman W (1996) Applied linear regression models. Irwin, Chicago
- Páez A, Uchida T, Miyamoto K (2002) A general framework for estimation and inference of geographically weighted regression models: 2. Spatial association and model specification tests. *Environ Plan A* 34(5):883–894
- Páez A, Long F, Farber S (2008) Moving window approaches for hedonic price estimation: an empirical comparison of modelling techniques. *Urban Stud* 45(8):1565–1581
- Spiegelhalter D, Thomas A, Best N, Lunn D (2003) WinBUGS users manual, version 1.4. MRC Biostatistics Unit, Cambridge, UK
- Thomas A, Best N, Lunn D, Arnold R, Spiegelhalter D (2004) GeoBUGS user manual, version 1.2. MRC Biostatistics Unit, Institute of Public Health, Cambridge, UK
- Waller L, Zhu L, Gotway C, Gorman D, Gruenewald P (2007) Quantifying geographic variations in associations between alcohol distribution and violence: a comparison of geographically weighted regression and spatially varying coefficient models. *Stoch Environ Res Risk Assess* 21(5):573–588
- Wheeler DC (2007) Diagnostic tools and a remedial method for collinearity in geographically weighted regression. *Environ Plan A* 39(10):2464–2481
- Wheeler DC (2009) Simultaneous coefficient penalization and model selection in geo-graphically weighted regression: the geographically weighted lasso. *Environ Plan A* 41:722–742
- Wheeler DC, Calder C (2007) An assessment of coefficient accuracy in linear regression models with spatially varying coefficients. *J Geograph Syst* 9(2):145–166
- Wheeler DC, Tiefelsdorf M (2005) Multicollinearity and correlation among local regression coefficients in geographically weighted regression. *J Geograph Syst* 7:161–187
- Wheeler DC, De Roos AJ, Cerhan JR, Morton LM, Severson RK, Cozen W, Ward MH (2011) Spatial-temporal cluster analysis of non-Hodgkin lymphoma in the NCI-SEER NHL study. *Environ Health* 10:63
- Wood S (2006) Generalized additive models: an introduction with R. Chapman and Hall/CRC, Boca Raton

Peter M. Atkinson and Christopher D. Lloyd

Contents

74.1	Introduction	1462
74.2	Characterizing Spatial Variation	1463
74.2.1	The Experimental Variogram	1463
74.2.2	Variogram Model Fitting	1464
74.2.3	Example	1466
74.3	Spatial Interpolation with Kriging	1468
74.3.1	Simple and Ordinary Kriging	1468
74.3.2	Simulation	1470
74.4	The Change of Support Problem	1470
74.4.1	Regularization	1470
74.4.2	Variogram Deconvolution	1471
74.4.3	Area-to-Point Kriging	1471
74.5	Geostatistics and Regional Science	1473
74.6	Conclusions	1474
	References	1474

Abstract

Characterizing the spatial structure of variables in the regional sciences is important for several reasons. Firstly, the spatial structure may itself be of interest. The structure of a population variable tells us something about how the population is configured spatially. For example, is the population clustered by some properties, but not others? Secondly, mapping variables from sparse

P.M. Atkinson (✉)

Geography and Environment, University of Southampton, Southampton, UK
e-mail: P.M.Atkinson@soton.ac.uk

C.D. Lloyd

School of Environmental Sciences, University of Liverpool, Liverpool, UK
e-mail: C.D.Lloyd@liverpool.ac.uk

sample observations or transferring values between areal units requires knowledge of how the property of interest varies spatially. Thirdly, we require knowledge of spatial variation in order to design sampling strategies which make the most of the effort, time, and money expended in sampling. Geostatistics comprises a set of principles and tools which can be applied to characterize or model spatial variation and use that model to optimize the mapping, simulation, and sampling of spatial properties. This chapter provides an introduction to some key ideas in geostatistics, with a particular focus on the kinds of applications which may be of interest for regional scientists.

74.1 Introduction

Geostatistics provides a body of techniques which can be used to characterize spatial variation, interpolate and simulate spatial variables, and design optimal spatial sampling strategies (Journel and Huijbregts 1978; Goovaerts 1997; Haining et al. 2010). To date, most applications of geostatistics have been in the physical sciences. However, there is an increasing diversity of research in the social sciences which makes use of geostatistical methods. This chapter reviews some of the key principles underlying geostatistics and considers the contributions these approaches may make in a regional science context.

Underlying classical geostatistics is the random function (RF) model (Journel and Huijbregts 1978; Chilès and Delfiner 1999). A RF is a set of random variables (RVs) which vary as a function of location \mathbf{x} . A RV is a stochastic process, a simple discrete example of which is the rolling of a die with the outcome being a value between one and six. Each outcome of this process is termed a *realization*. In most practical applications of geostatistics, variables are continuous. As an example, in a geostatistical framework, population density z at location \mathbf{x} is considered a RV, and the set of RVs at locations $z(\mathbf{x}_i)$, $i = 1, \dots, n$ is a RF, or more precisely, since n is finite, a random *vector*. Observations of population densities are termed regionalized variables (ReVs), and they are treated as stochastic realizations of an underlying RF.

In geostatistics, it is common to fit a stationary RF model (Journel and Huijbregts 1978; Chilès and Delfiner 1999). First, a spatially stationary (i.e., constant) mean parameter is usually defined. However, various alternatives are common in which the mean is allowed to vary across space (Goovaerts 1997) including those RF models that include a spatially varying “trend” (e.g., a two-dimensional polynomial or some spatially varying function of covariates; Hudson and Wackernagel 1994). A second parameter (more strictly a model comprising several parameters) that is usually defined is the stationary spatial covariance function (which implies second-order stationarity; see Journel and Huijbregts 1978) or variogram (defined further below, which implies intrinsic stationarity), either of which can be used to represent the “character” of spatial variation. Since the variogram function is more widely applicable than the spatial covariance function, we will refer to it here and in the remainder of the text. The mean and

variogram are, thus, the parameters that define the RF, and, thus, it is these parameters that need to be estimated to provide a useful model for geostatistical inference (e.g., prediction of the value of the property of interest at some unobserved location).

One might be wondering what utility the variogram brings and why stationarity is required as a modeling decision. We offer the following lay person explanation. In geostatistics, the central modeling decision is that the character of spatial variation can be treated as “stationary” (i.e., independent of location). This decision to fit a RF model that is parameterized by a stationary function representing the character of spatial variation is crucial to geostatistical inference. It means that it is possible to (i) lump together the observed spatial variation from different pairs of locations separated by specific distances, and possibly directions (this is essentially what the variogram does); (ii) use that information to infer the character of spatial variation at new pairs of locations; and (iii) crucially, given an observation representing one of a new pair of locations, predict at the other unobserved location based on the imputed character of spatial variation and, thus, how related we expect the two locations to be.

We now show (i) how the RF model can be estimated through calculation of the empirical or experimental variogram and the fitting of mathematical models with parameters and (ii) how the parameterized RF model can be used for spatial prediction through a technique known as Kriging and in other geostatistical operations.

74.2 Characterizing Spatial Variation

Most geostatistical analyses proceed with a summary of the properties of the data, and this is generally regarded as good practice. As in any statistical analysis, summary statistics and the histogram may suggest that the data should be transformed in some way (e.g., taking a logarithm so that the transformed distribution is approximately normal).

The spatial structure of a variable can be characterized with several structure functions, most commonly the variogram (outlined below).

74.2.1 The Experimental Variogram

The variogram cloud relates (half the) the squared differences (i.e., the semivariances) between paired data values to the distance (and possibly direction) by which they are separated, and it, thus, provides a summary of how different observations are as a function of distance.

The variogram (or, more properly, semivariogram) collapses the information contained in the variogram cloud; the semivariances within distance bands are averaged, and so, for example, there is an average semivariance for data pairs separated by 0–10 m, 10–20 m, and so on. The plot of average semivariance against

distance band indicates how spatially dependent the variable is. In practice, many properties tend to become more dissimilar as the separation distance increases.

The experimental variogram, $\hat{\gamma}(\mathbf{h})$, is estimated from $p(\mathbf{h})$ paired observations, $z(\mathbf{x}_\alpha), z(\mathbf{x}_\alpha + \mathbf{h}), \alpha = 1, 2, \dots, p(\mathbf{h})$ with:

$$\hat{\gamma}(\mathbf{h}) = \frac{1}{2p(\mathbf{h})} \sum_{\alpha=1}^{p(\mathbf{h})} \{z(\mathbf{x}_\alpha) - z(\mathbf{x}_\alpha + \mathbf{h})\}^2 \quad (74.1)$$

Semivariances can be computed within particular directional tolerances. For example, only data pairs which are aligned approximately north–south could be included in Eq. (74.1). In this way, it is possible to assess how spatial variation differs as a function of direction.

Where the variable of interest is a rate calculated from a numerator and a population-based denominator (e.g., as for disease rate), the standard experimental variogram can be modified easily to account for spatially varying underlying populations. That is, greater weight can be given to zones with large populations, and thus, it is possible to account for spatially varying uncertainties in rates (see Goovaerts 2005; Goovaerts et al. 2005).

74.2.2 Variogram Model Fitting

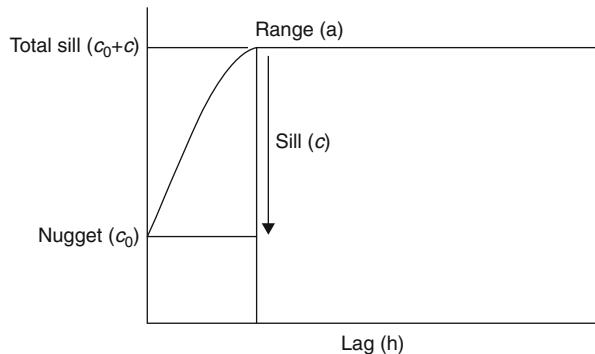
The variogram may be used as a means in itself to analyze the spatial structure of a variable (e.g., Berberoglu et al. 2000; Lloyd et al. 2004). More commonly, it is the first stage in Kriging-based spatial interpolation, as described below. A model can be fitted to the variogram, for example, by weighted least squares (Cressie 1985; McBratney and Webster 1986). Then the coefficients of the fitted model can be used as an input to the Kriging process. In principle, the variogram model provides information on how much weight is given to observations as a function of their distance from prediction locations. This use of information on spatial structure makes Kriging distinct from other commonly applied methods of interpolation such as inverse distance weighting, whereby weights are a simple function of distance and no account is taken of the specific spatial structure of the data being analyzed. The most commonly used models are taken from a set of “permissible” or “authorized” models.

Some of the most commonly used authorized models are detailed below. The nugget effect model, which indicates measurement error and microscale variation (variation at a distance smaller than the sample spacing), is given by:

$$\gamma(h) = \begin{cases} 0 & \text{for } h = 0 \\ c_0 & \text{for } |h| > 0 \end{cases} \quad (74.2)$$

Three of the most frequently used bounded models are the spherical model, the exponential model, and the Gaussian model, and these are each defined below. Each

Fig. 74.1 Bounded variogram model: nugget effect and spherical model



of these models is “bounded” in that each has a maximum value of semivariance defined by a “sill” parameter. The exponential model is given by:

$$\gamma(h) = c \left[1 - \exp\left(-\frac{h}{d}\right) \right] \quad (74.3)$$

where c is the sill of the exponential model and d is the non-linear distance parameter. The exponential model reaches the sill asymptotically, and the practical range is $3d$ (i.e., the separation at which approximately 95 % of the sill is reached).

The spherical model is a very widely used variogram model, and its form corresponds well with what is often observed in real-world studies, with an almost linear growth in semivariance to a particular separation and then stabilization (Armstrong 1998). It is given by:

$$\gamma(h) = \begin{cases} c \left[1.5 \frac{h}{a} - 0.5 \left(\frac{h}{a} \right)^3 \right] & \text{if } h \leq a \\ c & \text{if } h > a \end{cases} \quad (74.4)$$

where a is the non-linear parameter, known as the range.

The Gaussian model is given by:

$$\gamma(h) = c \left[1 - \exp\left(-\frac{h^2}{d^2}\right) \right] \quad (74.5)$$

As for the exponential model, the Gaussian model does not reach a sill at a finite distance, but rather approaches a defined sill asymptotically. The practical range is $a \times \sqrt{3}$ (Journel and Huijbregts 1978). Variograms with parabolic behavior at the origin, as represented by the Gaussian model here, are indicative of very regular spatial variation (Journel and Huijbregts 1978). Authorized models can be used in positive linear combination where a single model is insufficient to represent well the form of the variogram (Webster and McBratney 1989). Figure 74.1 depicts a variogram model comprising the combination of a nugget effect and a spherical component. In practice, the combination of the nugget effect and one or more permissible models is common.

Where the experimental variogram does not reach a maximum, it may be desirable to fit a trend model to the data and model the residuals with a bounded model. In some circumstances, it may be desirable to select an unbounded variogram model. The most widely used unbounded model is the power model:

$$\gamma(h) = mh^\omega \quad (74.6)$$

where ω is a power $0 < \omega < 2$ with a positive slope, m (Deutsch and Journel 1998); the linear model is a special case of the power model.

74.2.3 Example

The variogram characterizes the spatial structure in a variable. Where the variable represents some characteristic of a population, the variogram captures information on the dominant spatial scales of variation in that population property. Figure 74.3 shows variograms estimated from log-ratio transformed data on the percentage of Catholics in Northern Ireland in 1971, 1991, and 2001 for 1-km-square cells. The data are from the Census of Population, and the log ratios are computed with $1/\sqrt{2} \times \ln(\text{Catholics by religion } (\%)/\text{non-Catholics by religion } (\%)$). The percentages were computed with $n_1 + 1$ and $n_2 + 1$, where n_1 is the number of Catholics and n_2 is the number of non-Catholics for each 1-km cell; this avoids logging zeros and reflects uncertainties in small counts. The data and transformation rationale are described in Lloyd (2010). Briefly, where the variables used sum to a constant (e.g., percentages sum to 100), use of raw values may be problematic, and Aitchison (1986) suggests that use of log ratios is a suitable approach; most applications of such approaches are with respect to analysis of compositions with multiple parts. The present analysis is based on only one variable (religion or community background) with only two groups (Catholic or non-Catholic), which can, thus, be expressed as a single ratio. However, Filzmoser et al. (2009) show that univariate statistical methods should not be applied directly to (raw) compositional data. Furthermore, Pawlowsky and Burger (1992) argue that structural analysis should not be conducted using raw proportions or percentages as the constant-sum constraint may lead to a distorted picture of the spatial covariance structure and possibly erroneous interpretations.

The variograms in Fig. 74.2 were fitted with a nugget effect and two spherical components. This combination of models reflects nested spatial structures. That is, the variogram models exhibit two breaks corresponding to distances of approximately 5 and 30 km, and these spatial scales are represented by the two model ranges. The ranges and the nugget effects are similar for each of the 3 census years, while the total sill, which represents the magnitude of variation, for 1971 is much smaller than the sills for 1991 and 2001. These results suggest that the major scales of variation did not change between 1971 and 2001 – the areas which were dominated by Catholics or Protestants were broadly similar. What did change was the magnitude of differences between places. In support of previous studies

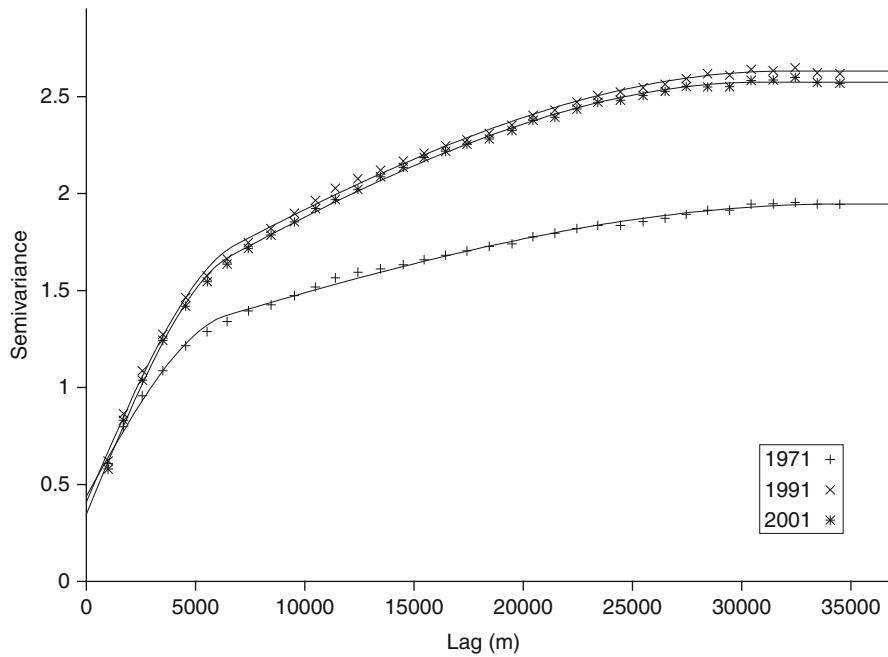


Fig. 74.2 Variogram of religion log ratios for 1971, 1991, and 2001 in Northern Ireland

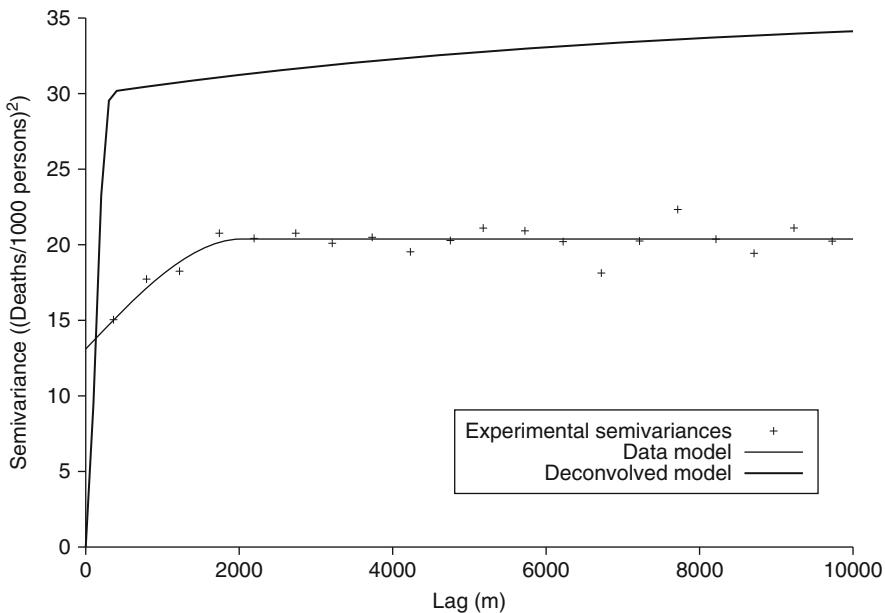


Fig. 74.3 Variogram of deaths/1,000 persons for SOAs with fitted model and deconvolved model

of residential segregation in Northern Ireland, the variograms suggest that the concentrations of Catholics and Protestants increased between 1971 and 1991, but there was little change (at the Northern Ireland (NI) scale at least) between 1991 and 2001. In other words, in 1991 Catholic areas were more Catholic than they had been in 1971, while Protestant areas had also become more Protestant. Lloyd (2012) discussed these results in more depth and also estimates variograms locally to enable assessment of how population spatial structures vary across NI. The derivation of local variograms is discussed by Lloyd (2011).

74.3 Spatial Interpolation with Kriging

74.3.1 Simple and Ordinary Kriging

Ordinary Kriging (OK) predictions are weighted averages of the n locally available data. The OK weights define the best linear unbiased predictor (BLUP). The OK prediction, $\hat{z}_{OK}(\mathbf{x}_0)$, is defined as:

$$\hat{z}_{OK}(\mathbf{x}_0) = \sum_{\alpha=1}^n \lambda_{\alpha}^{OK} z(\mathbf{x}_{\alpha}) \quad (74.7)$$

with the constraint that the weights, λ_{α}^{OK} , sum to one to ensure an unbiased prediction:

$$\sum_{\alpha=1}^n \lambda_{\alpha}^{OK} = 1 \quad (74.8)$$

Using the Kriging system, appropriate weights can be determined, and these are multiplied by the available observations before summing the products to obtain the predicted value. These weights are derived given the coefficients of a model fitted to the variogram (or another function such as the covariance function) (Oliver 2010).

The Kriging prediction error must have an expected value of zero:

$$E\{\hat{Z}_{OK}(\mathbf{x}_0) - Z(\mathbf{x}_0)\} = 0 \quad (74.9)$$

The Kriging (or prediction) variance, σ_{OK}^2 , is expressed as:

$$\begin{aligned} \hat{\sigma}_{OK}^2(\mathbf{x}_0) &= E[\{\hat{Z}_{OK}(\mathbf{x}_0) - Z(\mathbf{x}_0)\}^2] \\ &= 2 \sum_{\alpha=1}^n \lambda_{\alpha}^{OK} \gamma(\mathbf{x}_{\alpha} - \mathbf{x}_0) - \sum_{\alpha=1}^n \sum_{\beta=1}^n \lambda_{\alpha}^{OK} \lambda_{\beta}^{OK} \gamma(\mathbf{x}_{\alpha} - \mathbf{x}_{\beta}) \end{aligned} \quad (74.10)$$

That is, we seek the values of $\lambda_1, \dots, \lambda_n$ (the weights) that minimize this expression with the constraint that the weights sum to one [Eq. (74.8)].

This minimization is achieved through Lagrange multipliers. The conditions for the minimization are given by the OK system comprising $n + 1$ equations and $n + 1$ unknowns:

$$\left\{ \begin{array}{l} \sum_{\beta=1}^n \lambda_{\beta}^{\text{OK}} \gamma(\mathbf{x}_{\alpha} - \mathbf{x}_{\beta}) + \psi_{\text{OK}} = \gamma(\mathbf{x}_{\alpha} - \mathbf{x}_0) \alpha = 1, \dots, n \\ \sum_{\beta=1}^n \lambda_{\beta}^{\text{OK}} = 1 \end{array} \right. \quad (74.11)$$

where ψ_{OK} is a Lagrange multiplier. Given ψ_{OK} , the prediction variance of OK can be derived with:

$$\hat{\sigma}_{\text{OK}}^2 = \sum_{\alpha=1}^n \lambda_{\alpha}^{\text{OK}} \gamma(\mathbf{x}_{\alpha} - \mathbf{x}_0) + \psi_{\text{OK}} \quad (74.12)$$

The Kriging variance estimates the prediction variance (i.e., the square of the prediction error) based on the linear algebra given above. It is, thus, a measure of confidence in predictions and is a function of the form of the variogram, the sample configuration, and the sample support (Journel and Huijbregts 1978). The Kriging variance is, however, not conditional on the data values locally, and this has led some researchers to use alternative approaches such as conditional simulation (discussed in the next section) to build models of “spatial” uncertainty (Goovaerts 1997).

There are two standard varieties of OK: punctual OK and block OK. With punctual OK the predictions cover the same area or volume (the support, v ; see Atkinson and Tate 2000) as the observations. For example, if observations of some property of a population are made through a grid-based census with a grid cell size of 1 km, as in NI, then predictions made from those data will also have a support of (i.e., represent) 1-km cells. In block OK, the predictions are made to a larger support than the observations. With punctual OK the original observation data are honored. That is, they are retained in the output map. Block OK predictions are averages over areas (i.e., the support has increased). Thus, at \mathbf{x}_0 the prediction is not the same as an observation and does not need to honor it. In regional science contexts, data are rarely available on point supports, and often a concern may be how to predict from areas to points, a topic discussed below.

The choice of variogram model affects the Kriging weights and, therefore, the predictions. However, if the form of two models is similar at the origin of the variogram, then the two sets of results may be similar (Armstrong 1998). The choice of nugget effect may have marked implications for both the predictions and the Kriging variance. As the nugget effect is increased, the predictions become closer to the global average (Isaaks and Srivastava 1989).

Poisson Kriging provides an appropriate means to interpolate rare events such as mortality. The Poisson Kriging system incorporates the population-weighted mean of the rates (Goovaerts 2005) which accounts for the variability due to population

size, with larger weights where the population size is larger and, therefore, where the data may be considered more reliable.

Other forms of Kriging include Kriging with a trend model (where large-scale trends in the data are explicitly taken into account), cokriging (where information on secondary variables is used in the prediction process) (Wackernagel 2003), and indicator Kriging (where the data are transformed into a set of thresholds and the probability of exceeding thresholds is estimated) (Goovaerts, 1997).

74.3.2 Simulation

Kriging predictions are weighted moving averages, and thus, Kriging is a smoothing interpolator. In block Kriging, some amount of smoothing happens naturally through averaging over the larger support, but some occur due to the interpolation process itself, which effectively extends the support out to include the observations. The latter smoothing is an unwanted artifact of the interpolation process. A means of overcoming this unwanted smoothing is conditional simulation (Journel 1996; Goovaerts 1997; Dungan 1999). With conditional simulation, predictions are drawn from equally probable joint realizations of the RVs which comprise a RF model (Deutsch and Journel 1998). The simulated values are not the expected values (as in Kriging), but rather they are drawn from the conditional cumulative distribution function (ccdf), as a function of the available data, including both the observations and previously simulated data, and the (modeled) spatial variation. Since the approach is stochastic, multiple simulated realizations can be obtained. The range of values obtained represents the “spatial” uncertainty.

74.4 The Change of Support Problem

In regional science contexts, data are often available for zones rather than points. While individual or household level data are available in some contexts, these are usually provided without detailed spatial information, and spatially aggregated data usually offer the only means of exploring detailed spatial patterns. The data support, v , is defined as the geometrical size, shape, and orientation of the units associated with the measurements (Atkinson and Tate 2000). Thus, making predictions from areas to points corresponds to a change of support. Geostatistics offers the means to (i) explore how the spatial structure of a variable changes with change of support and (ii) change the support by interpolation to an alternative zonal system or to a quasi-point support (Schabenberger and Gotway 2005).

74.4.1 Regularization

For many applications, the variogram defined on a point support is not available. Indeed, it is physically impossible to measure on a point support given that all

measurements are integrals over a positive finite support. Thus, only values over a positive support (area) may be available. The variogram of such aggregated data is termed the regularized or areal variogram (Goovaerts 2008).

74.4.2 Variogram Deconvolution

Theoretically, given the point support variogram, it is possible to estimate the variogram for any support. Through variogram deconvolution, the point support variogram can be estimated from the variogram estimated from areal data. Atkinson and Curran (1995) derived the point support variogram from the areal support variogram, with the areas defined as regular grids, following an iterative procedure. Of greater relevance for many regional science applications is variogram deconvolution for irregular supports, that is, those cases where the data are available on irregular zones (such as census or administrative zones) rather than regular cells. Goovaerts (2008) presents a procedure for variogram deconvolution given irregular zones, and this method is implemented in the STIS software (see <http://www.biomedware.com/>).

The above deconvolution method is illustrated here through an example. The case study makes use of data on deaths per 1,000 of the population in Northern Ireland (NI). The death data are counts for 2008, and the denominator is given as the midyear estimates of total population in 2008 for super output areas (SOAs). Figure 74.3 gives the variogram estimated from deaths/1,000 persons over SOAs. The deconvolved model, derived using the method of Goovaerts (2008), is also given. The application of the deconvolved model for Kriging is illustrated below.

74.4.3 Area-to-Point Kriging

Given the deconvolution procedures outlined above, it is possible to make predictions at point locations given data defined on areal supports. Kyriakidis (2004) and Goovaerts (2008) show how the Kriging system is adapted in the case of areal data supports and point prediction locations.

Area-to-point Kriging is illustrated given the death rate data and the deconvolved variogram detailed in the previous section. Ordinary Kriging with Poisson population (2008 midyear estimates for each SOA) adjustment was applied with a population denominator of 1,000 (i.e., the deaths are rates per 1,000 of the population). The discretization geography was 1-km cells populated in 2001 (from 2001 Census of Population), with numbers of persons per cell (2001 Census counts) as weights. The destination geography (locations where estimates are required) was 1-km cells (as for the discretization geography). Figure 74.4 shows deaths/1,000 persons (A) for SOAs and (B) derived using area-to-point Kriging at 1-km cells. In areas covered by large SOAs (with lower density populations), the increased detail is particularly apparent.

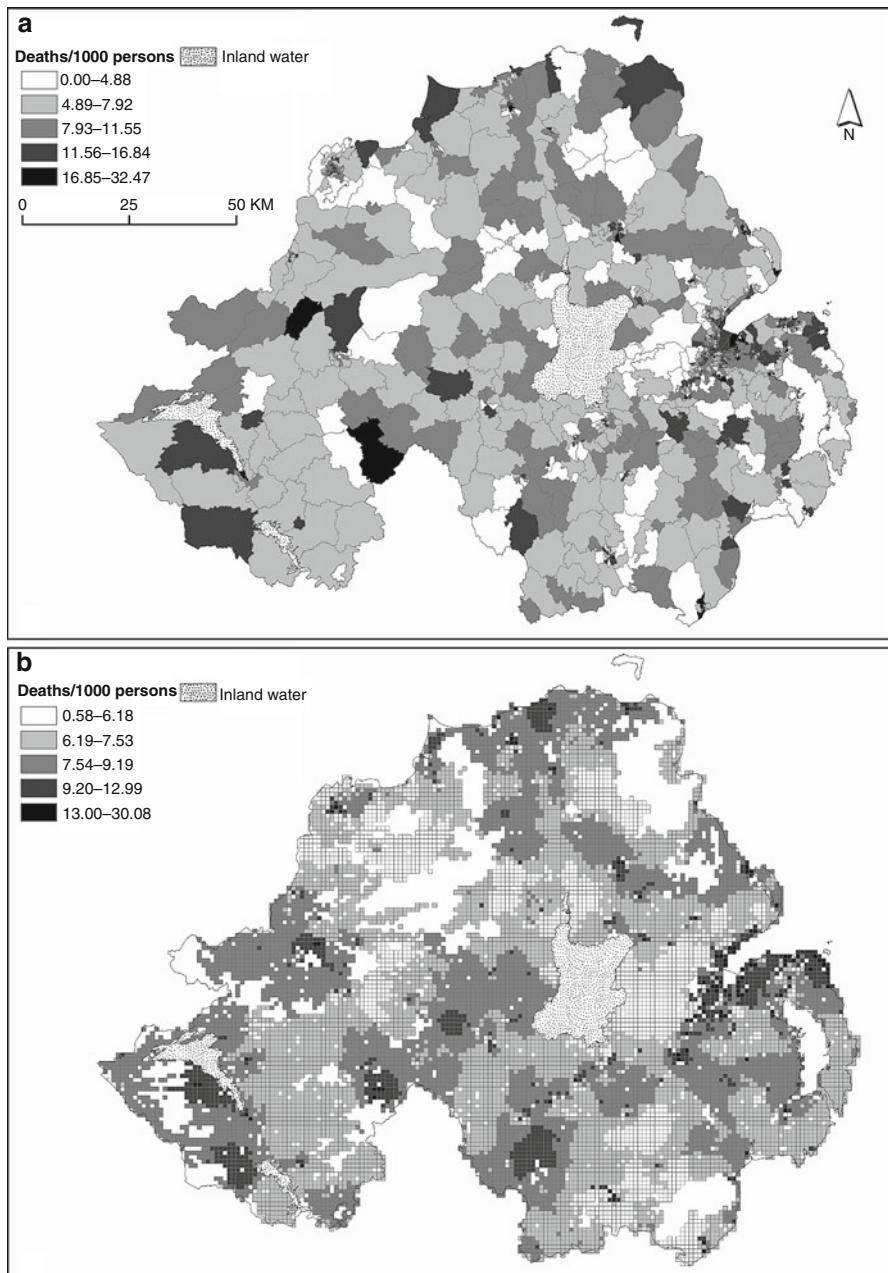


Fig. 74.4 Deaths/1,000 persons (a) for SOAs and (b) derived using area-to-point Kriging at 1-km cells (only populated cells shown). Source: 2001 Census, Output Area Boundaries. Crown copyright 2003

74.5 Geostatistics and Regional Science

Geostatistics originated in the mining industry through the work of Danie Krige in the 1950s and particularly through Georges Matheron who developed the Theory of Regionalized Variables in the 1960s and 1970s. From the original mining geology application, domain geostatistics found further application in the dominant field of petroleum geology, particularly through the Stanford Center for Reservoir Forecasting (SCRF) led by Andre Journel, but also in hydrogeology led by researchers such as Jaime Gómez-Hernández, in soil science led by the work of Richard Webster (see Webster and Oliver 2000) and Alex McBratney and subsequently Pierre Goovaerts, and in remote sensing and environmental science more generally. Thus, most applications of geostatistics can be divided into mining, petroleum, and “environmental”, and indeed, many conferences such as GeoENV organized themselves in this way for many years. However, the potential of geostatistics in the social sciences has been highlighted in recent research.

Much interest has been generated in the use of geostatistics to study a range of population-based properties, including rates based on census attributes as exemplified in this chapter and disease outcomes. This has been enabled by two key developments: (i) the pioneering work of Pascal Monestiez and Pierre Goovaerts on Poisson Kriging (Goovaerts 2005; Monestiez 2006), which enables the prediction of counts (e.g., mortality or morbidity) given an underlying population (e.g., at-risk population), and (ii) the parallel development in statistics of what has become known as model-based geostatistics. For example, in relation to the former, population-weighted variograms enable robust estimation in cases where the population underlying the observations is variable in magnitude. Moreover, Poisson Kriging allows the spatial interpolation of variables representing “rare” events, such as mortality.

Model-based geostatistics deserves special mention here as it represents an alternative Bayesian approach to geostatistics that has been made popular by statisticians such as Peter Diggle and is favored by the epidemiological community. Model-based geostatistics differs from the classical approach presented in this chapter in that the parameters of the RF model are regarded as uncertain, whereas in the classical approach these parameters are estimated through variogram model fitting and then regarded as “fixed” for the purposes of Kriging. As a result, model-based geostatistics conveys naturally the uncertainty in parameter estimation through to the prediction, resulting in a more complete description of prediction uncertainty. The general approach advocated by Diggle and others is two staged: (i) regression is used to predict the variable of interest from covariates, and (ii) geostatistics is used to predict the residuals at unobserved locations based on spatial dependence in those residuals. Since model-based geostatistics naturally allows for the adoption of a range of regression link functions (including the Poisson and logistic functions), the framework is general and can be applied readily to the population-based prediction problems described above. Thus, the development of new approaches, or adaptations of existing methods, has made the potential range of applications of geostatistical methods in regional science much wider.

The utility of area-to-point Kriging is much greater than may at first seem the case. Censuses are the primary means of enumerating the populations of the world's nations. However, the majority of them use arbitrary areal units to convey the original data, which were measured at the individual level, to the public (e.g., for reasons of preventing disclosure of personal information). Consequently, key variables relating to the world's population are obscured by (i) the aggregation effect which is effectively the convolution discussed above and (ii) the zonation effect which means that alternative realizations of the set of arbitrary boundaries may lead to very different realizations. This is the well-known modifiable areal unit problem (MAUP) which has dogged census analysis for decades. Area-to-point Kriging tackles the MAUP head on allowing researchers to bypass the sampling frame imposed by a particular set of areal units and re-present the census data on any support. While there are obvious limits to what can be achieved (it is not possible to generate more information than one started with), the results are visually stunning and do represent an optimal linear solution to the problem (Kyriakidis 2004; Goovaerts 2008). Additionally, applications that require common supports for multiple time periods (e.g., for which zonal systems may differ) may also benefit from recent developments in variogram deconvolution and area-to-point Kriging.

74.6 Conclusions

This chapter has introduced classical geostatistics and emphasized the advances that have been made in recent years that have opened up the possibility of applying geostatistics in social science and regional science. The ease of access to software which implements modern geostatistical methods means that researchers are increasingly likely to be held back only by their knowledge, rather than appropriate tools. It is hoped that this chapter will provide a useful starting point for those who wish to explore geostatistical methods in the context of regional science.

Acknowledgments The authors thank the anonymous referees for their comments and thank the editors for their patience while this chapter was finalized. The Northern Ireland Statistics and Research Agency are thanked for access to data. Census output is Crown copyright and is reproduced with the permission of the Controller of HMSO and the Queen's Printer for Scotland. Northern Ireland Statistics and Research Agency, 2001 Census: Standard Area Statistics (Northern Ireland) [computer file]. ESRC/JISC Census Programme, Census Dissemination Unit, Mimas (University of Manchester).

References

- Aitchison J (1986) The statistical analysis of compositional data. Chapman and Hall, London
Armstrong M (1998) Basic linear geostatistics. Springer, Berlin
Atkinson PM, Curran PJ (1995) Defining an optimal size of support for remote sensing investigations. IEEE Trans Geosci Remote Sens 33(3):768–776

- Atkinson PM, Tate NJ (2000) Spatial scale problems and geostatistical solutions: a review. *Prof Geogr* 52(4):607–623
- Berberoglu S, Lloyd CD, Atkinson PM, Curran PJ (2000) The integration of spectral and textural information using neural networks for land cover mapping in the Mediterranean. *Comput Geosci* 26(4):385–396
- Chilès J-P, Delfiner P (1999) *Geostatistics: modeling uncertainty*. Wiley, New York
- Cressie NAC (1985) Fitting variogram models by weighted least squares. *Math Geol* 17(5):563–586
- Deutsch CV, Journel AG (1998) *GSLIB: geostatistical software and user's guide*, 2nd edn. Oxford University Press, New York
- Dungan JL (1999) Conditional simulation. In: Stein A, van der Meer F, Gorte B (eds) *Spatial statistics for remote sensing*. Kluwer, Dordrecht, pp 135–152
- Filzmoser P, Hron K, Reimann C (2009) Univariate statistical analysis of environmental (compositional) data: problems and possibilities. *Sci Total Environ* 407(23):6100–6108
- Goovaerts P (1997) *Geostatistics for natural resources evaluation*. Oxford University Press, New York
- Goovaerts P (2005) Geostatistical analysis of disease data: estimation of cancer mortality risk from empirical frequencies using Poisson kriging. *Int J Health Geogr* 4:31pp
- Goovaerts P (2008) Kriging and semivariogram deconvolution in the presence of irregular geographical units. *Math Geosci* 40(1):101–128
- Goovaerts P, Jacquez GM, Greiling D (2005) Exploring scale-dependent correlations between cancer mortality rates using factorial kriging and population-weighted semivariograms. *Geogr Anal* 37(2):152–182
- Haining RP, Kerry R, Oliver MA (2010) Geography, spatial data analysis, and geostatistics: an overview. *Geogr Anal* 42(1):7–31
- Hudson G, Wackernagel H (1994) Mapping temperature using kriging with external drift: theory and an example from Scotland. *Int J Climatol* 14(1):77–91
- Isaaks EH, Srivastava RM (1989) *An introduction to applied geostatistics*. Oxford University Press, New York
- Journel AG (1996) Modelling uncertainty and spatial dependence: stochastic imaging. *Int J Geogr Inf Syst* 10(5):517–522
- Journel AG, Huijbregts CJ (1978) *Mining geostatistics*. Academic, London
- Kyriakidis PC (2004) A geostatistical framework for area-to-point spatial interpolation. *Geogr Anal* 36(3):259–289
- Lloyd CD (2010) Exploring population spatial concentrations in Northern Ireland by community background and other characteristics: an application of geographically weighted spatial statistics. *Int J Geogr Inf Sci* 24(8):1193–1221
- Lloyd CD (2011) *Local models for spatial analysis*, 2nd edn. CRC Press, Boca Raton
- Lloyd CD (2012) Analysing the spatial scale of population concentrations by religion in Northern Ireland using global and local variograms. *Int J Geogr Inf Sci* 26(1):57–73
- Lloyd CD, Berberoglu S, Curran PJ, Atkinson PM (2004) A comparison of texture measures for the per-field classification of Mediterranean land cover. *Int J Remote Sens* 25(19):3943–3965
- McBratney AB, Webster R (1986) Choosing functions for semi-variograms of soil properties and fitting them to sampling estimates. *J Soil Sci* 37(4):617–639
- Monestiez P, Dubroca L, Bonnin E, Durbec J-P, Guinet C (2006) Geostatistical modelling of spatial distribution of *Balaenoptera physalus* in the Northwestern Mediterranean Sea from sparse count data and heterogeneous observation efforts. *Ecol Model* 193(3–4):615–628
- Oliver MA (2010) The variogram and kriging. In: Fischer MM, Getis A (eds) *Handbook of applied spatial analysis. Software tools, methods and applications*. Springer, Berlin/Heidelberg/New York, pp 319–352
- Pawlowsky V, Burger H (1992) Spatial structure analysis of regionalized compositions. *Math Geol* 24(6):675–691

-
- Schabenberger O, Gotway CA (2005) Statistical methods for spatial data analysis. Chapman and Hall/CRC, Boca Raton
- Wackernagel H (2003) Multivariate geostatistics. An introduction with applications, 3rd edn. Springer, Berlin
- Webster R, McBratney AB (1989) On the Akaike information criterion for choosing models for variograms of soil properties. *J Soil Sci* 40(3):493–496
- Webster R, Oliver MA (2000) Geostatistics for environmental scientists. Wiley, Chichester

Daniel Griffith and Yongwan Chun

Contents

75.1	Introduction	1478
75.1.1	Spatial Autocorrelation: A Conceptual Overview	1478
75.1.2	Spatial Filtering: A Conceptual Overview	1480
75.2	Eigenvector Spatial Filtering	1481
75.2.1	Eigenvectors of a Spatial Weight Matrix	1481
75.2.2	An Eigenvector Spatial Filter Specification of the Linear Regression Model	1483
75.2.3	The Numerical Calculation of Eigenvectors	1486
75.3	Statistical Features of ESF	1488
75.4	Comparisons with the Spatial Lag Term	1496
75.5	An Empirical Example: An ESF Methodology Example	1499
75.6	Extensions to Spatial Interaction Data Analysis	1500
75.7	Conclusions	1502
Appendix A	1505	
References	1506	

Abstract

This chapter provides an introductory discussion of spatial autocorrelation (SA), which refers to correlation existing and observed in geospatial data, and which characterizes data values that are not independent, but rather are tied together in overlapping subsets within a given geographic landscape. This chapter summarizes the various interpretations of SA, one being map pattern. SA can be quantified in a number of different ways, too, one being with the Moran Coefficient. Spatial filtering is a statistical method whose goal is to obtain

D. Griffith (✉) • Y. Chun

Geospatial Information Sciences, School of Economic, Political and Policy Sciences, University of Texas at Dallas, Richardson, TX, USA

e-mail: dagriffith@utdallas.edu; ywchun@utdallas.edu

enhanced and robust results in a spatial data analysis by decomposing a spatial variable into trend, a spatially structured random component (i.e., spatial stochastic signal), and random noise. Its aim is to separate spatially structured random components from both trend and random noise, and, consequently, leads statistical modeling to sounder statistical inference and useful visualization. This separation procedure can involve eigenfunctions of the matrix version of the numerator of the Moran Coefficient. This chapter summarizes the eigenvector spatial filtering (ESF) conceptual material, and presents the computer code for implementing ESF in R, Matlab, MINITAB, FORTRAN, and SAS. Next, it demonstrates that eigenvector spatial filter estimators are unbiased, efficient, and consistent. Finally, it summarizes an ESF empirical example application, and the extension of ESF to spatial interaction modeling.

75.1 Introduction

Spatial autocorrelation (SA) refers to correlation existing and observed in geospatial data – data collected together with their locational position tags on a two-dimensional (2-D) surface. This concept describes data values that are not independent but rather are tied together in overlapping subsets within a given geographic landscape. Relative location, commonly expressed in terms of geographic closeness, partly determines SA. Tobler's (1969, p. 7) first law of geography provides a generalized view of how phenomena occur across space within this context: "Everything is related to everything else, but near things are more related than distant things." This description indicates that a data value at a given location tends to be (dis)similar to those values for the same variable at nearby locations. For example, house price and house value comparisons among effectively equivalent neighboring houses partly govern a real estate market. That is, real-world phenomena are not likely to occur in a random fashion but rather in a spatially structured manner, over space.

Griffith (1992) points out that, in practice, spatial scientists essentially interpret SA in one of nine different ways: self-correlation, map pattern, a diagnostic tool, a missing variables surrogate, a spatial process mechanism, a spatial spillover effect, an outcome of areal unit demarcation (re. the MAUP), redundant information, and a nuisance parameter. Discussion in this chapter exploits the map pattern interpretation of SA.

75.1.1 Spatial Autocorrelation: A Conceptual Overview

SA (Getis 2010) can be interpreted in different ways. First, it can be interpreted literally as self-correlation which arises in 2-D space. Unlike the conventional Pearson's product moment correlation coefficient that measures co-variability of paired values in two variables, SA measures correlation among paired values in

Table 75.1 The 1981 Cliff-Ord example, with an extension

Model specification	(Pseudo-)R ²	Residual spatial autocorrelation z-score
$Y = \alpha + \beta X + \varepsilon$	0.404	1.87
$Y = \alpha X^\beta e^\varepsilon$	0.517	1.41
$Y = \alpha + \beta \text{LN}(X - \delta) + \varepsilon$	0.735	0.76

a single variable based on relative spatial locations. Because it focuses on a tendency among values of a variable based on their spatial closeness, SA is measured within the combinatorial context of all possible pairs of observed values for a given variable where corresponding weights that are determined by spatial closeness identify the pairings of interest. Second, SA can be interpreted as map pattern. In regional science, an analysis often involves datasets of individual observations post-stratified by geographical units such as census blocks/block groups/tracts, county boundaries, and country borders. The choropleth mapping of a variable using such areal units portrays a pattern over space. A tendency for similarity or dissimilarity for neighboring values on such a map can be directly interpreted as SA. While large clusters of similar values on a map indicate positive SA, an alternating map pattern, in which the tendency is for values to be dissimilar to those of their neighbors, constitutes negative SA. Third, SA can be interpreted as a diagnostic tool for misspecification in statistical modeling such as linear regression analysis. For example, detected SA among residuals in such a regression analysis can suggest that the linear relationship specified between a response variable and a covariate needs to be replaced with a nonlinear functional description of the relationship. Cliff and Ord (1981, pp. 209) furnish a useful illustration of this situation by employing an empirical example in which the 1961 population as a percentage of the 1926 population (Y) by Eire county ($n = 26$) is cast as a function of arterial road network accessibility (X). Table 75.1 summarizes selected results from their analysis, supplemented with an extension. As the nonlinear relationship between Y and X is better articulated, SA displayed by the corresponding regression residuals decreases to a clearly statistically nonsignificant level. Finally, SA also can imply that variables are missing from a regression model specification, in which case it serves as a surrogate for unexplained variation in the response variable in question. When a variable is missing from a model specification that, indeed, accounts for a substantial amount of variation in a dependent variable, then the unexplained variation may occur in the form of SA among residuals. If a missing variable forms an underlying map pattern – that is, significant SA exists – then residuals are very likely to exhibit a conspicuous level of SA.

Research in the spatial sciences in general, and in geography and regional science in particular, has addressed the numerical quantification of SA. The Moran coefficient (MC) is the most commonly used quantitative measure of SA; it is analogous to the Pearson's correlation coefficient. The MC may be written as

$$\frac{\sum_{i=1}^n \sum_{j=1}^n c_{ij}(y_i - \bar{y})(y_j - \bar{y}) / \sum_{i=1}^n \sum_{j=1}^n c_{ij}}{\sqrt{\sum_{i=1}^n (y_i - \bar{y})^2 / n} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2 / n}} = \frac{n}{\sum_{i=1}^n \sum_{j=1}^n c_{ij}} \frac{\sum_{i=1}^n (y_i - \bar{y}) \left[\sum_{j=1}^n c_{ij}(y_j - \bar{y}) \right]}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

where c_{ij} is a non-negative nonzero spatial weight tying together the pair of areal units i and j , n is the number of areal units, and \bar{y} is the arithmetic mean of response variable Y . The value of c_{ij} reflects spatial closeness between areal units i and j . The numerator of the MC contains the pairs of values $(y_i - \bar{y})$ and $\sum_{j=1}^n c_{ij}(y_j - \bar{y})$, whose

graphic portrayal is the Moran scatterplot. In most cases, the MC utilizes spatial connectivity information for a set of areal units. That is, in binary form, $c_{ij} = 1$ when areal units i and j have a common boundary, and 0 otherwise; the rook's case definition of spatial closeness is when all common boundaries have nonzero length (i.e., they are lines), whereas the queen's case definition is when some also have zero length boundaries (i.e., they are points). The range of the MC is slightly different from that of Pearson's correlation coefficient, which covers the interval $[-1, 1]$. Mathematical quantities of the n -by- n matrix of c_{ij} values, C , called eigenvalues determine the exact range of a MC (de Jong et al. 1984). Often, the largest value is slightly larger than 1, and the smallest value is between -1 and -0.5 .

Statistical hypothesis testing involving the MC can be based upon a normal distribution assumption; Cliff and Ord (1981) derive its sampling distribution theory as such. The expected value for the MC is $-1/(n-1)$ under the null hypothesis of zero SA. This expected value is the threshold value separating positive and negative SA. That is, while an MC value that is greater than this expected value indicates positive SA, one that is less than this expected value indicates negative SA. Although equations for the standard deviation of the MC are rather complicated (see Cliff and Ord 1981 for details), it can be well approximated by $\sqrt{\frac{2}{\sum_{i=1}^n \sum_{j=1}^n c_{ij}}}$ when $n \geq 20$ (Griffith 2010); the summation term, $\sum_{i=1}^n \sum_{j=1}^n c_{ij}$,

counts the number of 1s in matrix C , which equals twice the number of neighbors in a surface partition (i.e., a link from areal unit i to areal unit j plus a link from areal unit j to areal unit i). A z-score calculated for the MC with the foregoing expected value and standard deviation furnishes a basis for statistical inference that asymptotically is the standard normal distribution.

75.1.2 Spatial Filtering: A Conceptual Overview

Spatial filtering (Tiefelsdorf and Griffith 2007) is a statistical method whose goal is to obtain enhanced and robust results in a spatial data analysis. The fundamental idea is based upon a decomposition of a spatial variable into the

following three components: trend, spatially structured random component (i.e., spatial stochastic signal), and random noise. Its aim is to separate spatially structured random components from both trend and random noise, and, consequently, leads statistical modeling to sounder statistical inference and useful visualization. This separation procedure involves sophisticated mathematical operators and produces the aforementioned eigenvalues as well as their accompanying eigenvectors (the couplings are called eigenfunctions).

A limited number of spatial filtering methods currently exist, including autoregressive linear operators, Getis's G_i -based specification, interpoint distance matrix eigenfunctions, and spatial weight matrix eigenfunctions. These methods utilize different constructed mathematical operators to decompose a spatial variable into the three components. First, following the Cochrane-Orcutt procedure for time series analysis, autoregressive linear operators can be used to prewhiten dependent variable to render independent and identically distributed (i.i.d.) random errors (Tobler 1975; Haining 1991). This spatial filtering method takes the matrix form $(\mathbf{I} - \rho\mathbf{C})$, where ρ is a SA parameter, and \mathbf{I} is an n-by-n identity matrix. With the parameter ρ estimated solely from a variable Y , a prewhitened dependent variable is generated by $(\mathbf{I} - \rho\mathbf{C})Y$. Second, Getis's G_i -based method converts each spatially correlated variable into two variates, one capturing SA and the other containing nonspatial systematic and random effects (Getis 1990). Regression with spatial and nonspatial variates enables a separation of SA components from trend and white noise components. Third, Borcard and Legendre (2002) propose a procedure called principal coordinates of neighbor matrices (PCNM) that exploits eigenfunctions of the n-by-n matrix of truncated geographic distances among locations [truncation can be at the (effective) range of SA as identified by, say, a semivariogram]. They use eigenvectors corresponding to positive eigenvalues as spatial descriptors in regression or canonical analysis. They argue that this method can be applied to any set of locations providing a good coverage of a given geographic landscape. Fourth, eigenvector spatial filtering (ESF) methodology utilizes eigenvectors extracted from spatial weight matrix C (Griffith 2003). Details of this latter methodology are discussed in the next sections because this chapter focuses on the ESF method.

Spatial filtering methods are increasingly utilized in regression settings, although they can be applied to individual variables, too. Especially, the ESF methodology has been utilized in different model specifications, including linear regression (Getis and Griffith 2002), generalized linear models (Chun 2008; Griffith 2004), and space-time panel models (Chun and Griffith 2011; Patuelli et al. 2011).

75.2 Eigenvector Spatial Filtering

ESF uses a set of synthetic proxy variables, which are extracted as eigenvectors from a spatial weight matrix C that ties geographic objects together in space and then adds these vectors as control variables to a model specification. These control variables identify and isolate the stochastic spatial dependencies among the georeferenced observations, thus allowing model building to proceed as if the observations are independent.

75.2.1 Eigenvectors of a Spatial Weight Matrix

The ESF method utilizes a mathematical decomposition, whose results are called eigenfunctions, of the following transformed spatial weight matrix:

$$(\mathbf{I} - \mathbf{1}\mathbf{1}^T/n)\mathbf{C}(\mathbf{I} - \mathbf{1}\mathbf{1}^T/n)$$

where $\mathbf{1}$ is an n-by-1 vector of ones, and T denotes the matrix transpose operator. The decomposition generates n eigenvectors and their associated n eigenvalues. In descending order, the n eigenvalues can be denoted as $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \lambda_3, \dots, \lambda_n)$. That is, the eigenvalues range between the largest eigenvalue that is positive, λ_1 , and the smallest eigenvalue that is negative, λ_n . The corresponding n eigenvectors can be denoted as $\mathbf{E} = (\mathbf{E}_1, \mathbf{E}_2, \mathbf{E}_3, \dots, \mathbf{E}_n)$, where each eigenvector, \mathbf{E}_j , is an n-by-1 vector. In matrix notation, the decomposition can be expressed as

$$\mathbf{M}\mathbf{C}\mathbf{M} = \mathbf{E} \boldsymbol{\Lambda} \mathbf{E}^T \quad (75.1)$$

where $\mathbf{M} = (\mathbf{I} - \mathbf{1}\mathbf{1}^T/n)$, the projection matrix that centers a variable and that commonly appears in standard multivariate statistical theory, and $\boldsymbol{\Lambda}$ is an n-by-n diagonal matrix whose diagonal elements are the set of n eigenvalues ($\lambda_1, \lambda_2, \lambda_3, \dots, \lambda_n$).

The eigenfunctions have some important properties. First, the eigenvectors are mutually orthogonal and uncorrelated (Griffith 2000): the symmetry of matrix \mathbf{C} ensures orthogonality, and the projection matrix \mathbf{M} ensures eigenvectors with zero means, guaranteeing uncorrelatedness. That is, $\mathbf{E}\mathbf{E}^T = \mathbf{I}$ and $\mathbf{E}^T\mathbf{1} = \mathbf{0}$, and the correlation between any pair of eigenvectors, say \mathbf{E}_i and \mathbf{E}_j , is zero when $i \neq j$. Second, the eigenvectors portray distinct, selected map patterns. Tiefelsdorf and Boots (1995) show that each eigenvector portrays a different map pattern exhibiting a specified level of SA when it is mapped onto the n areal units associated with the corresponding spatial weight matrix \mathbf{C} . They also show that the MC value for a mapped eigenvector is equal to a function of its corresponding eigenvalue (i.e., $MC_j = \frac{n}{\mathbf{1}^T \mathbf{C} \mathbf{1}} \cdot \lambda_j$, for \mathbf{E}_j). Third, given a spatial weight matrix, the feasible range of MC values is determined by the largest and smallest eigenvalues; that is, by λ_1 and λ_n (de Jong et al. 1984). Based upon these properties, the eigenvectors can be interpreted as follows (Griffith 2003):

The first eigenvector, \mathbf{E}_1 , is the set of real numbers that has the largest MC value achievable by any set of real numbers for the spatial arrangement defined by the spatial weight matrix \mathbf{C} ; the second eigenvector, \mathbf{E}_2 , is the set of real numbers that has the largest achievable MC value by any set that is uncorrelated with \mathbf{E}_1 ; the third eigenvector, \mathbf{E}_3 , is the set of real numbers that has the largest achievable MC value by any set that is uncorrelated with both \mathbf{E}_1 and \mathbf{E}_2 ; the fourth eigenvector is the fourth such set of values; and so on through \mathbf{E}_n , the set of real numbers that has the largest negative MC value achievable by any set that is uncorrelated with the preceding $(n-1)$ eigenvectors. As such, these eigenvectors furnish distinct map

pattern descriptions of latent SA in spatial variables, because they are mutually both orthogonal and uncorrelated.

[Figure 75.1](#) presents selected eigenvectors generated with the rook's definition of adjacency used to construct a 0–1 binary spatial weight matrix for the 73 municipalities of Puerto Rico. [Figure 75.1a](#) portrays the boundaries of the 73 municipalities together with their neighboring structure: the red lines indicate that c_{ij} for a pair of connected municipalities equals 1; otherwise, it equals 0. [Figure 75.1b](#) is a map of the first eigenvector, \mathbf{E}_1 , which portrays the maximum positive SA possible for the established matrix \mathbf{C} . Its MC value is 1.0938. This map pattern shows two large clusters of areal units, one of positive numbers and one of negative numbers. [Figure 75.1c](#) portrays the map pattern of \mathbf{E}_{18} , whose MC value is 0.2832. It still indicates positive SA but with a degree much weaker than that for eigenvector \mathbf{E}_1 . It depicts smaller-sized areal unit clusters of similar values. In contrast, [Fig. 75.1d](#) is the map of eigenvector \mathbf{E}_{73} , which portrays the extreme negative SA possible with this spatial tessellation. Its pattern is alternating positive and negative values, and its MC is -0.5657 .

Puerto Rico's east–west elongation coupled with its irregular municipality surface partitioning somewhat obscures the conspicuousness of map patterns portrayed by the eigenvectors of matrix \mathbf{C} . This notion is more clearly depicted by the eigenvectors of a regular square tessellation. Consider a landscape comprising 900 pixels forming a 30-by-30 square tessellation ([Fig. 75.2](#)). [Figure 75.2a](#) portrays the boundaries of the 900 pixels together with their neighboring structure. [Figure 75.2b](#) portrays a global SA pattern: it contains a northwest-southeast swath crossing a northeast-southwest swath of similar values – with a relative MC value of 0.9922. [Figure 75.2c](#) portrays a regional SA pattern: it contains 36 moderate-sized clusters of similar values – with a relative MC value of 0.8314. [Figure 75.2d](#) portrays a more local SA pattern: it contains 100 small clusters of similar values – with a relative MC value of 0.5358. In all three cases, appropriate spatial aggregation would yield negative SA (i.e., negative MC values).

75.2.2 An Eigenvector Spatial Filter Specification of the Linear Regression Model

ESF methodology utilizes the eigenvectors calculated in the previous section. In detail, it accounts for SA with a linear combination of the eigenvectors. In linear regression, the ESF model specification may be written as

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta}_X + \mathbf{E}_k\boldsymbol{\beta}_k + \boldsymbol{\epsilon}$$

where \mathbf{X} is an n -by- $(p+1)$ matrix containing covariates, $\boldsymbol{\beta}_X$ is the corresponding $(p+1)$ -by-1 vector of regression parameters, \mathbf{E}_k is an n -by- k matrix containing k eigenvectors, $\boldsymbol{\beta}_k$ is the corresponding vector of regression parameters, and $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \mathbf{I}\sigma^2)$ is an n -by-1 error vector whose elements are i.i.d normal random variates. Because the linear combination of the eigenvectors, $\mathbf{E}_k\boldsymbol{\beta}_k$, accounts for

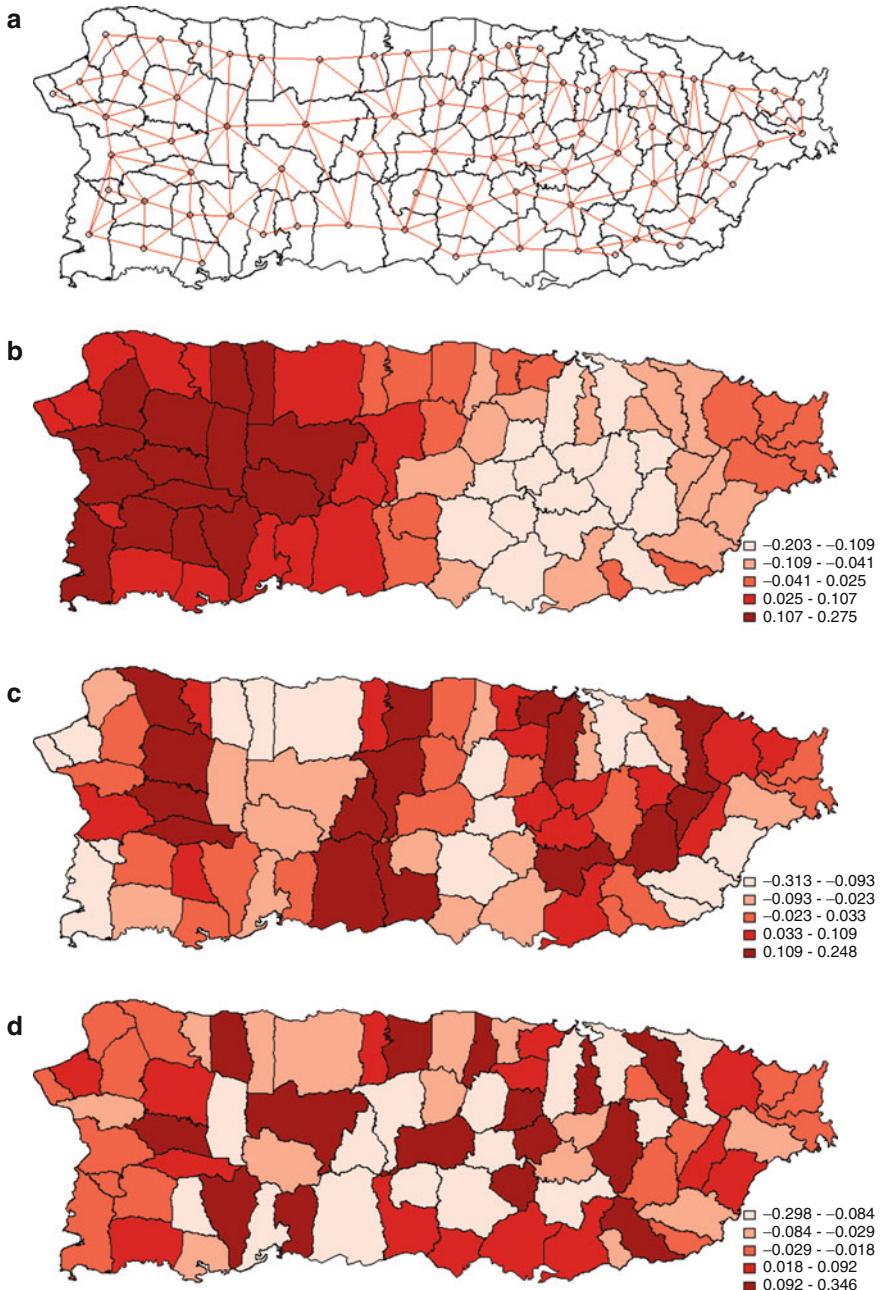


Fig. 75.1 The maps of selected eigenvectors based upon matrix C constructed for the municipalities of Puerto Rico: (a) the boundaries and links of the 73 municipalities, (b) the first eigenvector, E_1 , (c) the eighteenth eigenvector, E_{18} , and (d) the nth eigenvector, E_n

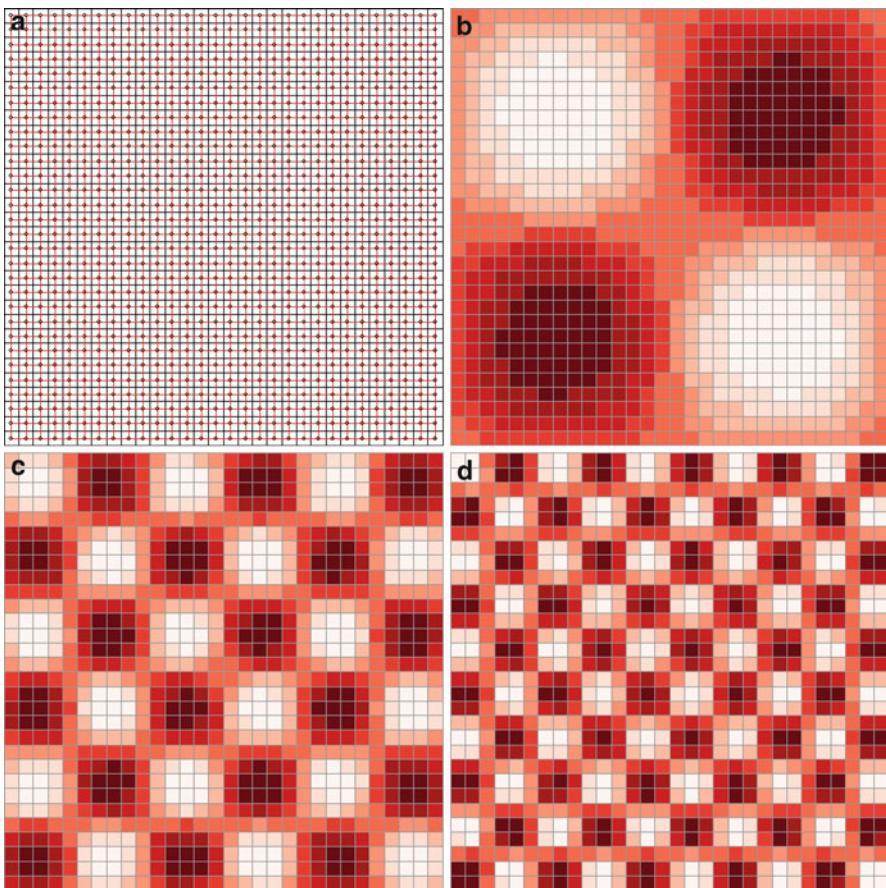


Fig. 75.2 The maps of selected eigenvectors based upon matrix \mathbf{C} constructed for the pixels of a 30-by-30 square tessellation: (a) the boundaries and links of the 900 pixels, (b) eigenvector $\mathbf{E}_{2,2}$, (c) eigenvector $\mathbf{E}_{6,6}$, and (d) eigenvector $\mathbf{E}_{10,10}$

SA, the ESF linear regression specification does not suffer from SA in its residuals. Addition of the eigenvectors in the regression equation does not change the expected conditional mean of Y because the mean of each eigenvector is zero.

Initializing ESF methodology requires the identification of a feasible set of eigenvectors. This procedure involves two steps. In the first step, a candidate set of eigenvectors, which is a noticeably smaller subset (i.e., $K \ll n$) of the entire set of eigenvectors, can be demarcated based upon several criteria. Eigenvectors whose MC values are close to the expected value of MC do not explain much spatial variation. These eigenvectors can be eliminated from a candidate set. Also, one of the eigenvalues of \mathbf{MCM} is zero due to the rank deficiency of matrix \mathbf{M} ; it can be eliminated, but the model specification actually already includes it through vector $\mathbf{1}$.

attached to the mean response. The candidate set can be further restricted to only eigenvectors portraying positive SA, if the MC for a response variable displays positive SA, or to only eigenvectors portraying negative SA, if the MC for the response variable displays negative SA; most empirical datasets display positive SA. Considering these factors, one criterion for identifying this set is a minimum MC of 0.25, which relates to roughly 5% of the variance in a response variable being attributable to SA (see Appendix A).

In the second step, a smaller set of eigenvectors can be identified from its candidate set using a stepwise regression selection technique. One way to select an eigenvector is to maximize model fit at each step through statistical significance (e.g., invoking a 10%, 5%, or 1% level). This selection can be implemented easily with conventional stepwise regression procedures. Another way to select an eigenvector is to minimize residual SA in each stepwise regression iteration. This selection procedure can be repeated until a prespecified level of residual SA (e.g., $MC \approx 0$) is achieved.

75.2.3 The Numerical Calculation of Eigenvectors

Spatial science practitioners find eigenfunction calculation one of the most, if not the most, difficult tasks in implementing the ESF methodology. Most statistical software packages provide functions to calculate eigenvectors. Extracting eigenvectors from a spatial weight matrix can be executed with the following three steps: (1) loading (or creating) a spatial weight matrix \mathbf{C} , (2) transforming the spatial weight matrix to its **MCM** version in Eq. (75.1), and (3) computing eigenvectors with software-specific functions or commands. This section outlines a wide array of computer codes for completing this task.

75.2.3.1 R Code for Generating MCM Eigenvectors

The *spdep* package of R provides functions to handle spatial datasets and to conduct spatial data analysis. The package can be loaded with the `library()` function:

```
> library(spdep)
```

For the first step, the spatial neighbor structure needs to be loaded from an existing file (e.g., *PuertoRico.gal*). A .GAL file can be created with GeoDa software. The function `read.gal()` reads a .GAL file and stores its neighbor structure in the .nb object format of *spdep*:

```
> pr.nb <- read.gal("PuertoRico.gal")
```

Although an .nb object contains lists of spatial neighbors, a listw object contains lists of actual weight values. A spatial weight matrix can be created directly from a listw object. The `nb2listw()` function requires an argument for a spatial weight matrix type. In the following code line, a spatial weight matrix is created from a listw object created as a binary type (`style="B"`), with the `listw2mat()` function:

```
> pr.listb <- nb2listw(pr.nb, style="B")
> B <- listw2mat(pr.listb)
```

In the second step, the matrix **MCM** can be created with matrix operations. The `length()` function returns the number of observations in the `pr.nb` object:

```
> n <- length(pr.nb)
> M <- diag(n) - matrix(1,n,n)/n
> MBM <- M %*% B %*% M
```

Finally, the function `eigen()` generates eigenvalues and eigenvectors. In the following code, `eig$vectors` contains n-by-n eigenvectors, and `eig$values` contain n eigenvalues. The second line selects only eigenvectors with $MC_j/MC_1 > 0.25$. The third line changes the column names of the EV objects:

```
> eig <- eigen(MBM, symmetric=T)
> EV <- as.data.frame(eig$vectors[,eig$values/
eig$values[1]>0.25])
> colnames(EV) <- paste("EV", 1:NCOL(EV), sep="")
```

The EV objects can be further utilized in a regression analysis in R, or can be exported in a different file format, such as CSV, for other purposes.

75.2.3.2 Matlab Code for Generating MCM Eigenvectors

Matlab does not have a built-in function to read a .GAL file. But the spatial neighbor structure in a .GAL file can be imported with Matlab functions to open and read a file. The [Computer Code 1](#) window provides sample Matlab code to generate eigenvectors.

The first five lines of code (code block ①) open the PuertoRico.gal file in order to store all file contents in the `lines` variable and then close the file. The next four lines of code (code block ②) retrieve the number of observations, `n`, which is stored in the first line of a .GAL file, and then create an n-by-n matrix of all zeros for a binary spatial weight matrix, `B`. The `for` statement code block (code block ③) updates the matrix `B` by assigning 1s for spatial neighbors from the PuertoRico.gal file. The final six lines of code (code block ④) generate eigenvalues and eigenvectors from matrix **MCM** and then select only eigenvectors with $MC_j/MC_1 > 0.25$.

75.2.3.3 MINITAB Code for Generating MCM Eigenvectors

MINITAB supports matrix operations and works with the 0–1 binary spatial weight matrix. The [Computer Code 2](#) window provides sample MINITAB code for computing eigenvectors.

The first block of code reads in the spatial weight matrix. The next block of code constructs matrix **M**. The third block of code computes **MCM**. The fourth block of code calculates the eigenvalues and their associated eigenvectors. The fifth block of code converts the eigenvalues to MCs. The final block of code identifies for which eigenfunctions $MC_j/MC_1 > 0.25$.

75.2.3.4 FORTRAN Code for Generating MCM Eigenvectors

The IMSL subroutines package supports reliable matrix operations for FORTRAN. The [Computer Code 3](#) window provides sample FORTAN code for computing eigenvectors.

```

filename = 'PuertoRico.gal';
fID = fopen(filename,'r');
gal = textscan(fID, '%s', 'delimiter', '\n', ' whitespace', '');
fclose(fID);
lines = gal{1};

header = lines{1};
headerNum = sscanf(header, '%f');
n = headerNum(2);
B = zeros(n);

for i = 1:n
    idline = sscanf(lines{i*2}, '%f');
    nbs = sscanf(lines{i*2+1}, '%f');

    nbloc = zeros(length(nbs),2);
    B(idline(1),nbs) = 1;
end

M = eye(n) - ones(n)/n;
MBM = M*B*M;
[evec, eval]=eig(MBM);
eval=diag(eval);
sel = eval./eval(1)>0.25;
EV = evec(:,sel);

```

Computer Code 1 Sample Matlab code for generating eigenvectors from a spatial weight matrix

The first part of the code reads in a spatial neighbors' file, similar to a .GAL file; its structure is id, number of neighbors, and list of neighbor ids. The first line of this file contains n. The next block of code constructs matrix **MCM**. Subroutine **DEVCSF** calculates the eigenvalues and their associated eigenvectors. The final block of code identifies for which eigenfunctions $MC_j/MC_1 > 0.25$.

75.2.3.5 SAS Code for Generating MCM Eigenvectors

The IML language facilitates eigenfunction calculations by SAS. The [Computer Code 4](#) window provides sample SAS code for computing eigenvectors.

As with the preceding FORTRAN code, SAS works with the spatial weight matrix as input. The first part of the code constructs matrix **M**. IML imports both matrix **C** and matrix **M**. It employs matrix operations to construction **MCM** and subroutine EIGEN to compute the eigenfunctions. Finally, it outputs a working file of eigenvalues (STEP3) and working file of eigenvectors (STEP4).

75.3 Statistical Features of ESF

Construction of an ESF begins with determination of a candidate set of eigenvectors, which is a substantially smaller subset of the total set of n eigenvectors. One criterion for identifying this set is a minimum MC of 0.25, which relates to roughly 5% of the variance in a response variable being attributable to SA (see Appendix A). Another

```

READ 73 73 M1;
FILE 'C:\PR-CONN.TXT';
FORMAT(2(5X,30F2.0,/,5X,13F2.0).
LET K1=73
SET C1
K1(1)
END

COPY C1 M2
TRANS M2 M3
MULT M2 M3 M2
LET K3=-1/K1
MULT K3 M2 M2
DIAG C1 M3
ADD M3 M2 M2

MULT M2 M1 M3
MULT M3 M2 M3

EIGEN M1 C1
EIGEN M3 C2 M2

SET C4
K1(1)
END
MULT M1 C4 C4
LET C3=K1*C2/SUM(C4)
LET C4=C3/MAX(C3)
SET C5
1:K1
END

COPY C5 C6;
USE C4 = 0.25:1.
LET K10=K1+100
COPY M2 C101-CK10
PRINT C6
END

```

} (2.1)

} (2.2)

} (2.3)

} (2.4)

} (2.5)

} (2.6)

Computer Code 2 Sample MINITAB code for generating eigenvectors from a spatial weight matrix

criterion devised by Chun and Griffith (2009) builds on the level of SA detected in a response variable, Y, and may be summarized as follows:

$$MC_j \geq 2.9970 - \frac{2.8805}{1 + e^{-0.6606 - 0.2525 z_{MC}}} \quad (75.2)$$

where z_{MC} denotes the z-score of the MC for the response variable Y (or some transformed version of it, such as a Box-Cox power transformation, if used). Two weaknesses of ESF construction concern the selection of eigenvectors from a candidate set to include in the constructed variate and the increase in number of candidate and selected eigenvectors with increasing n.

To date, eigenvector selection has been undertaken with stepwise regression procedures, an approach plagued by the pretesting problem. One redeeming feature of the eigenvectors is that they are both orthogonal and uncorrelated.

```

USE MSIMSL
    IMPLICIT DOUBLE PRECISION (A-H,O-Z)
    INTEGER INEIGH(3000,75),NI(3000)
    DOUBLE PRECISION C(3000,3000),W(3000,3000),ID(3000),EVAL(3000),MC,
    MCMAX

    OPEN(5,FILE='C:\INPUT.NEI')
    READ(5,*) N
    SUMC=0.0D0
    DO 5 I=1,N
    READ(5,*) ID(I),NI(I),(INEIGH(I,J),J=1,NI(I))
    5 SUMC=SUMC+REAL(NI(I))
    DO 20 I=1,N
    DO 18 J=1,N
    18 C(I,J) = 0.0D0
    DO 19 J=1,NI(I)
    19 C(I,INEIGH(I,J))=1.0D0
    20 CONTINUE

    DO 40 I=1,N
    DO 39 J=1,N
    39 W(I,J) = C(I,J) - REAL(NI(I))/REAL(N) - REAL(NI(J))/REAL(N)
    C
    + SUMC/REAL(N**2)
    40 CONTINUE

    49 CALL DEVCSF(N,W,3000,EVAL,C,3000)

    MCMAX=0.0D0
    DO 50 I=1,N
    MC=(REAL(N)/SUMC)*EVAL(I)
    IF(MC.GE.MCMAX) MCMAX=MC
    50 WRITE(7,1000) ID(I),EVAL(I),MC
    K=0
    DO 60 J=1,N
    IF((REAL(N)/SUMC)*EVAL(J)/MCMAX.GT.0.25D0) GOTO 60
    WRITE(6,*) J
    K=K+1
    DO 59 I=1,N
    W(I,K)=C(I,J)
    59 CONTINUE
    60 CONTINUE
    WRITE(6,*) 'NUMBER OF PROMINENT EIGENVECTORS = ',K

```

Computer Code 3 Sample FORTRAN code for generating eigenvectors from a spatial weight matrix

Accordingly, a Bonferroni adjustment, for example, furnishes the appropriate significance level to use with a stepwise procedure. The number of eigenvectors in a candidate set determines this adjustment. A simulation experiment employing i.i.d. random normal data illustrates this contention: the selection probability used is $p = 0.01$; the estimated distribution of selected eigenvectors is exactly that for a binomial distribution with $p = 0.01$ (see Tables 75.2 and 75.3). The Bonferroni adjustment – $0.01/(\# \text{ candidate eigenvectors})$ – more or less fully corrects for excess vector selection. In addition, for reasonably sized samples, essentially only 1 vector might be selected by chance, and its percentage of variance accounted for in the response variable tends to be trivial.

Using the Frisch-Waugh-Lovell theorem, Pace et al. (2011) confirm that ESF produces unbiased estimates of covariate regression parameters for data generated with a spatial error (i.e., SAR) model specification. They also find that for synthetic

```

DATA STEP2;
  ARRAY M{73} M1-M73;
  DO I=1 TO 73;
  DO J=1 TO 73;
    M{J} = -1/73;
    IF I=J THEN M{J}=1-1/73;
  END;
  OUTPUT;
END;
DROP I J;
RUN;

PROC IML;
USE STEP1; READ ALL INTO C;
USE STEP2; READ ALL INTO M;
IDEN=I(70);
CM=C*M;
MCM=M*CM;

CALL EIGEN(EVALS,EVECS,MCM);

CREATE STEP3 FROM EVALS;
APPEND FROM EVALS;
CREATE STEP4 FROM EVECS;
APPEND FROM EVECS;
QUIT;

```

Computer Code 4 Sample SAS code for generating eigenvectors from a spatial weight matrix

data generated with a spatial lag model specification, an ESF specification tends to result in some bias. ESF yields unbiased covariate parameter estimates if SA is present in the covariates, but not in the response variable. If SA is present in the response variable, then a large number of eigenvectors may be needed to ensure unbiased parameter estimates. Nevertheless, they conclude that ESF reduces bias found in an ordinary least squares (OLS) solution, thus improving upon it.

To illustrate unbiasedness, consider the generalized least squares estimator (GLS) – which also is the maximum likelihood estimator – of the constant mean (i.e., no covariates are present in a model specification) SAR model specification, which is given by $\hat{\mu} = \mathbf{1}^T \mathbf{V} \mathbf{Y} / \mathbf{1}^T \mathbf{V} \mathbf{1}$. Its expected value is

$$E[\mathbf{1}^T \mathbf{V} \mathbf{Y} / \mathbf{1}^T \mathbf{V} \mathbf{1}] = \mathbf{1}^T \mathbf{V} E[\mathbf{Y}] / \mathbf{1}^T \mathbf{V} \mathbf{1} = \mathbf{1}^T \mathbf{V} E[\mu \mathbf{1} + \mathbf{V}^{-1/2} \boldsymbol{\varepsilon}] / \mathbf{1}^T \mathbf{V} \mathbf{1} = \mu$$

where $\boldsymbol{\varepsilon} \sim N(0, \sigma_\varepsilon^2)$, \mathbf{V} is the inverse covariance matrix [e.g., $(\mathbf{I} - \rho \mathbf{W})^T (\mathbf{I} - \rho \mathbf{W})$ for the SAR model, where \mathbf{W} is a row-standardized spatial weight matrix; $\mathbf{W} = \mathbf{D}^{-1} \mathbf{C}$, and where $d_{ii} = \sum_{j=1}^n c_{ij}$, and $d_{ij} = 0$ for $i \neq j$]. Meanwhile, the ESF estimator of the constant mean model specification is given by $\mathbf{1}^T \mathbf{Y} / n$. Its expected value is as follows:

$$E[\mathbf{1}^T \mathbf{Y} / n] = \mathbf{1}^T E[\mathbf{Y}] / n = \mathbf{1}^T E[\mu \mathbf{1} + \mathbf{E}_k \mathbf{E}_k^T \mathbf{Y} + \mathbf{V}^{-1/2} \boldsymbol{\varepsilon}] / n = \mu$$

Table 75.3 Summary results of Bonferroni adjusted eigenvector selection simulation experiments, with 10,000 replications

P-by-P	MC MC ₁	Binomial model estimation			Bonferroni adjustment			Maximum R ²	Maximum # selected vectors
		# candidate vectors	\hat{p}	Pearson chi-square deviance	Selection p	Frequency of no selected eigenvectors			
6-by-6	0.25	12	0.01007	1.0605	0.00083	9,894	0.31	2	2
	0.50	7	0.01007	1.0652	0.00143	9,882	0.29		
10-by-10	0.75	3	0.01043	1.0307	0.00333	9,893	0.28	2	2
	0.25	31	0.01122	1.0819	0.00032	9,901	0.10		
20-by-20	0.50	18	0.01067	1.0274	0.00056	9,894	0.14	1	1
	0.75	9	0.01057	1.0082	0.00111	9,906	0.09		
30-by-30	0.25	123	0.01103	1.0800	0.00008	9,895	0.04	2	2
	0.50	74	0.01062	1.0683	0.00014	9,914	0.04		
	0.75	32	0.01061	1.0132	0.00031	9,888	0.03	1	1
	0.25	276	0.01097	1.1050	0.00004	9,894	0.01		
	0.50	163	0.01059	1.0503	0.00006	9,892	0.01	1	1
	0.75	76	0.01006	1.0112	0.00013	9,879	0.01		

where \mathbf{E}_k are the K selected eigenvectors extracted from a binary 0–1 spatial weight matrix used to construct an eigenvector spatial filter and $\mathbf{1}^T \mathbf{E}_k = 0$ by construction. In other words, both estimators are unbiased.

The sampling variance of μ for the OLS estimator is $\sigma_e^2 \mathbf{1}^T \mathbf{V}^{-1} \mathbf{1} / n^2$, and for the preceding GLS estimator is $\sigma_e^2 / \mathbf{1}^T \mathbf{V} \mathbf{1}$. Meanwhile, the unbiased ESF estimator of σ_e^2 is given by

$$\{(n - K) / [\text{TR}(\mathbf{V}^{-1}) - \mathbf{E}_k^T \mathbf{V}^{-1} \mathbf{E}_k]\} s_{\text{ESF}}^2$$

where s_{ESF}^2 is the MSE for ESF. In other words, s_{ESF}^2 underestimates σ_e^2 . The sampling variance of μ for the OLS estimator is σ_e^2 / n . The relative efficiencies of the OLS, GLS, and ESF estimators of a constant mean based upon unbiased variance estimators are given by

$$\text{GLS: OLS} = n^2 / [\mathbf{1}^T \mathbf{V} \mathbf{1} \mathbf{1}^T \mathbf{V}^{-1} \mathbf{1}] \quad (75.3a)$$

$$\text{ESF: GLS} = (1 - \rho)^2 \quad (75.3b)$$

$$\text{ESF: OLS} = n / \mathbf{1}^T \mathbf{V}^{-1} \mathbf{1} \quad (75.3c)$$

Eq. (75.3a)–(75.3c) equals 1 when no SA is present (i.e., $\mathbf{V} = \mathbf{I}$).

A numerical exercise was executed comparing efficiencies employing a 30-by-30 square lattice, and a simulation experiment was conducted employing a 10-by-10 square lattice, both utilizing the rook's definition of geographic adjacency and a constant mean SAR model specification. The simulation employed normally distributed errors; all simulated pseudorandom numbers were converted to z-scores for each sample, the 31 eigenvectors for which $MC_j/MC_1 > 0.25$ and the 44 eigenvectors for which $MC_j/MC_1 > 0$, $\alpha = 0.01/31 \approx 0.00032$ or $0.10/44 \approx 0.00227$ or 0.10, and 10,000 replications.

Figure 75.3 portrays the efficiency results of the mean estimators [i.e., Eq. (75.3a)–(75.3c) values] for 11 selected values of ρ spanning the full range of positive SA. As reported by Cordy and Griffith (1993), the GLS estimator's relative efficiency vis-à-vis the OLS estimator is very modest, and reaches only 0.82% when $\rho = 0.9999$. The ESF estimator increasingly is more efficient than either the GLS, SAR, or OLS estimators, with these latter estimators' efficiencies decreasing quite precipitously with increasing ρ . As an aside, the tabulated simulation results in Fig. 75.3 almost perfectly align with their Eq. (75.3b) counterparts appearing in the Figure's graph.

One implication suggested by the tabular part of Fig. 75.3 is that efficiency does not improve very much by increasing the number of eigenvectors available for constructing an ESF. In other words, once the prominent eigenvectors are selected

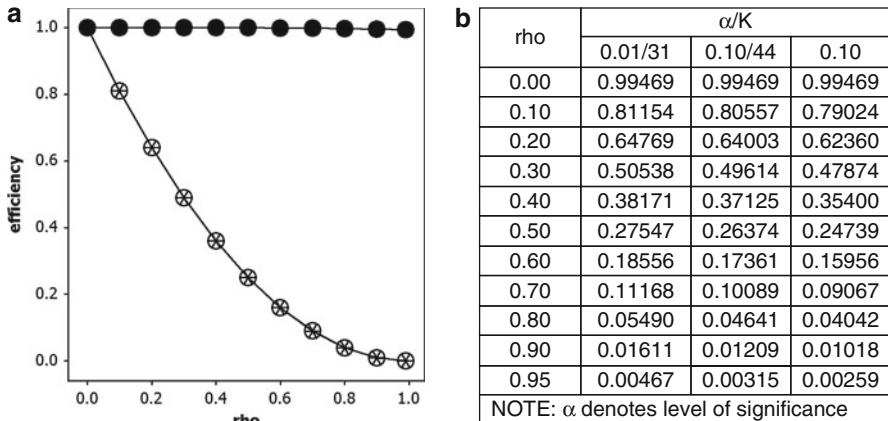


Fig. 75.3 Relative efficiency results for the mean. (a) the unbiased estimator efficiencies for a 30-by-30 regular square lattice: solid circle denotes $s^2_{\text{GLS}}/s^2_{\text{OLS}}$, open circle denotes $s^2_{\text{ESF}}/s^2_{\text{GLS}}$, and asterisk denotes $s^2_{\text{ESF}}/s^2_{\text{OLS}}$. (b) the biased ESF versus SAR mean estimate for a 10-by-10 regular square lattice

in an ESF construction, marginal vectors add little to the efficiency of the ESF estimator. This result supports the use of the preceding Bonferroni adjustment during the construction of an ESF.

Linear regression parameter estimates are consistent when the covariates are orthogonal and uncorrelated with the error term in a model specification. But the context for this property to hold is that n increases while the number of covariates remains unchanged. ESF tends to have an increasing number of eigenvectors with increasing n (as well as with increasing SA). Based on a selected set of data analyses, where n ranges from 49 to 2,379 (see Table 75.4), the following logistic equation describes this increase:

$$n_{\text{selected}} \approx \frac{n_{\text{candidate}}}{1 + 3.16749 e^{-0.11409z_{\text{MC}} + 0.00255n}} \quad (75.4)$$

where n_{selected} is the number of selected eigenvectors for constructing a filter, $n_{\text{candidate}}$ is the number of candidate eigenvectors for a given surface partitioning (e.g., those with a $MC_j/MC_1 > 0.25$), and z_{MC} is the z-score for the response variable's residual MC. Even when $n_{\text{candidate}} = n$ and for a completely connected planar graph, the limit of Eq. (75.4) divided by n (i.e., n_{selected}/n) is 0 as n goes to infinity. This result is in keeping with findings reported by Portnoy (1984). In other words, the covariate parameter estimates for an ESF model specification are consistent.

In summary, ESF estimators appear to be unbiased, efficient, and consistent. More comprehensive simulation experiments need to be completed to verify this generalization.

Table 75.4 Summary spatial autocorrelation and ESF results for selected geographic landscapes

Landscape	Attribute	MC	n	n _{selected}
Columbus, OH, census tracts	Crime	0.5206	49	6
North Carolina counties	SIDS	0.7091	100	10
Murray superfund site Thiessen polygons	Arsenic	0.2277	253	17
Mercer-Hall 25-by-20 agricultural field plots	Yield	0.4055	500	54
Toronto enumeration districts (1986)	Population density	0.6402	731	72
30-by-30 remotely sensed image pixels	Biomass	0.8679	900	198
Wiebe 125-by-12 agricultural field plots	Yield	0.7100	1,500	131
China counties	Population density	0.6908	2,379	181

75.4 Comparisons with the Spatial Lag Term

Conceptualization of the ESF model specification is based upon a spectral decomposition of a spatial weight matrix. The standard SAR specification furnishes the following approximation:

$$\mathbf{Y} = \rho \mathbf{WY} + (1 - \rho) \mu \mathbf{1} + \boldsymbol{\epsilon}$$

$$\mathbf{Y} = \rho \mathbf{D}^{-1} \mathbf{CY} + (1 - \rho) \mu \mathbf{1} + \boldsymbol{\epsilon}$$

where \mathbf{D} is a diagonal matrix whose cell (i,i) entry is the ith row sum of matrix \mathbf{C} ,

$$\mathbf{Y} = \rho \mathbf{D}^{-1} \mathbf{E} \Lambda \mathbf{E}^T \mathbf{Y} + (1 - \rho) \mu \mathbf{1} + \boldsymbol{\epsilon}$$

where \mathbf{E} is the n-by-n matrix of eigenvectors and Λ is the corresponding diagonal matrix of eigenvalues

$$\Rightarrow \mathbf{Y} = \mathbf{E} \mathbf{E}^T \mathbf{Y} + \mu \mathbf{1} + \boldsymbol{\xi}$$

This approximation requires the principal eigenfunction of matrix \mathbf{C} to be removed from consideration; in theory, vector $\mathbf{1}$ attached to the mean parameter μ replaces it. Furthermore, the remaining n-1 eigenvectors are asymptotically uncorrelated (see Griffith 2000).

A more direct ESF model specification is based upon the spatial lag model specification:

$$\mathbf{Y} = \mu \mathbf{1} + (\mathbf{I} - \rho \mathbf{C})^{-1} \boldsymbol{\epsilon}$$

$$\mathbf{MY} = \mu \mathbf{M1} + \mathbf{M}(\mathbf{I} - \rho \mathbf{C})^{-1} \boldsymbol{\epsilon}$$

where $\mathbf{M} = (\mathbf{I} - \mathbf{1}\mathbf{1}^T/n)$ is the standard projection matrix used to center a vector of data values,

$$\mathbf{M}(\mathbf{I} - \rho \mathbf{C})\mathbf{MY} = \mathbf{0} + \boldsymbol{\varepsilon}$$

$$\mathbf{MY} = \rho \mathbf{MCMY} + \boldsymbol{\varepsilon}$$

where matrix \mathbf{MCM} appears in the numerator of the MC,

$$\mathbf{Y} = \rho \mathbf{MCMY} + \bar{y}\mathbf{1} + \boldsymbol{\varepsilon}$$

$$\Rightarrow \mathbf{Y} = \mathbf{EE}^T\mathbf{Y} + \mu\mathbf{1} + \boldsymbol{\xi}$$

In other words, an ESF model specification approximates a spatial autoregressive model specification by approximating the included spatial lag variable. Removing those eigenvectors for which $\mathbf{E}_j^T\mathbf{Y} \approx 0$ reduces the covariate matrix \mathbf{E} used in a substitution for vector \mathbf{WY} from n-by-n to n-by-K, resulting in the general specification

$$\mathbf{Y} = \mathbf{E}_k\boldsymbol{\beta} + \mu\mathbf{1} + \boldsymbol{\xi}$$

where $\boldsymbol{\beta}$ is a K-by-1 vector of regression parameters. This particular general specification has the regression parameter estimates $\hat{\boldsymbol{\beta}} = (\mathbf{E}_k^T\mathbf{E}_k)^{-1}\mathbf{E}_k^T\mathbf{Y} = \mathbf{E}_k^T\mathbf{Y}$, yielding

$$\mathbf{Y} = \mathbf{E}_k\mathbf{E}_k^T\mathbf{Y} + \mu\mathbf{1} + \boldsymbol{\xi}$$

This reduction in matrix dimension can be achieved with stepwise regression techniques, which tremendously benefit from the mutual orthogonality and uncorrelatedness of the eigenvectors of modified matrix \mathbf{MCM} (two properties asymptotically achieved for matrix \mathbf{C} itself). From standard linear regression theory, the individual parameter estimate variances are given by the diagonal elements of matrix

$$\left[\begin{pmatrix} \mathbf{1}^T \\ \mathbf{E}_k^T \end{pmatrix} \begin{pmatrix} \mathbf{1} \\ \mathbf{E}_k \end{pmatrix} \right]^{-1} \sigma_\varepsilon^2 = \begin{pmatrix} \frac{1}{n} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_k \end{pmatrix} \sigma_\varepsilon^2$$

which furnishes the variance term used in the preceding estimator efficiency assessment.

The preceding discussion raises the question asking how well a set of selected eigenvectors approximates the spatial lag variate \mathbf{WY} . Figure 75.4 portrays the prediction of \mathbf{WY} by judiciously selected eigenvectors for the eight-specimen geospatial datasets used to assess consistency (see Table 75.4). An ESF description of vector \mathbf{Y} for these examples requires between 7% and 22% of the total number of eigenvectors. In contrast, an ESF description of vector \mathbf{WY} for these examples requires between 9% and 24% of the total number of eigenvectors, only a very slight increase in the number needed for describing variable \mathbf{Y} . In this context, Eq. (75.4) becomes

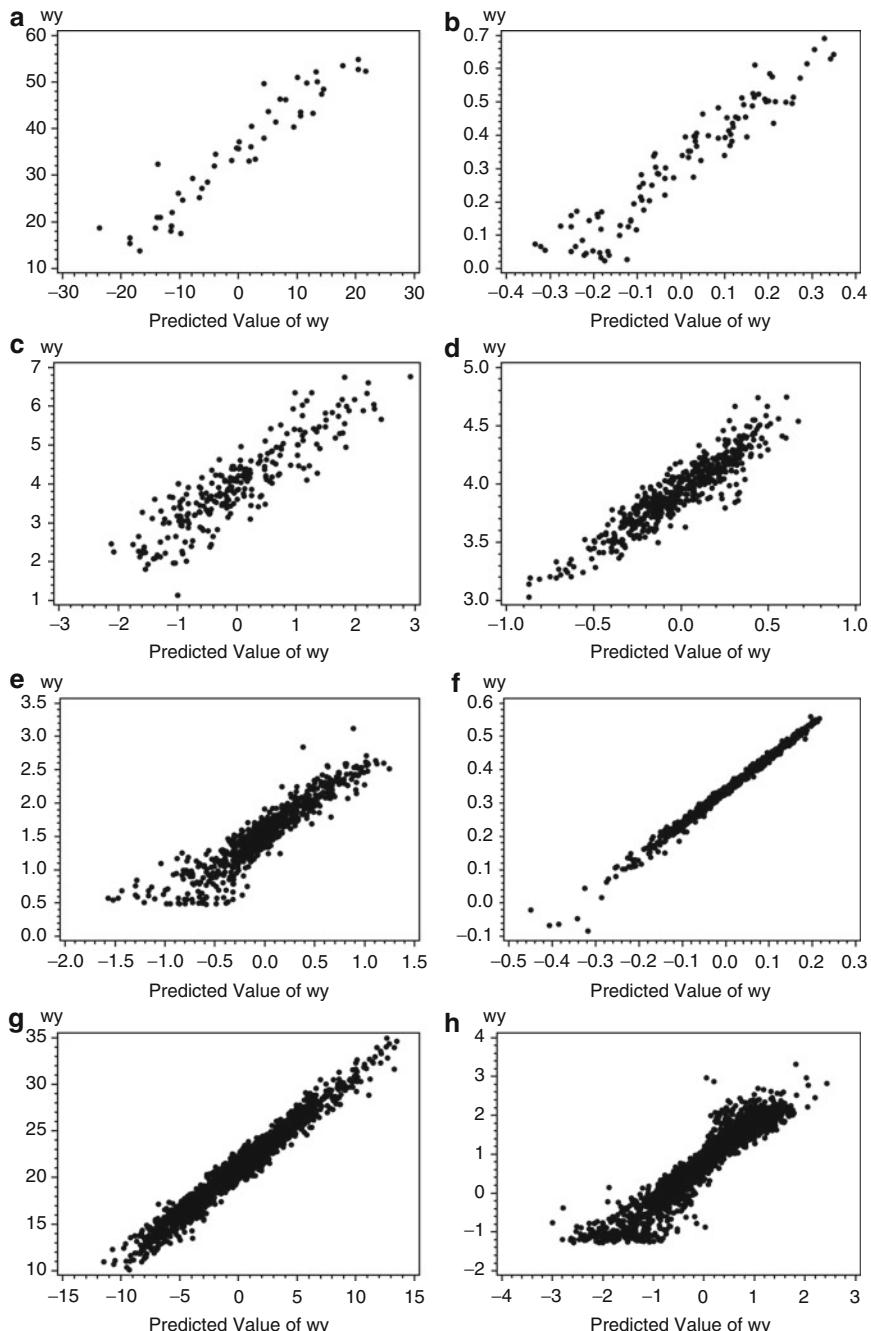


Fig. 75.4 Scatterplots of observed versus ESF-predicted WY vectors for the geographic landscapes listed in Table 75.4

$$n_{\text{selected}} \approx \frac{n_{\text{candidate}}}{1 + 0.85432 e^{-0.06967z_{\text{MC}} + 0.00189n}}$$

which still asymptotically goes to 0 when divided by n.

In summary, ESF furnishes a dimension reduction substitution for the spatial lag variate \mathbf{WY} in a spatial statistical model specification. Because the eigenvectors are fixed for a given matrix \mathbf{C} , this dimension reduction greatly simplifies accounting for SA in regression models.

75.5 An Empirical Example: An ESF Methodology Example

The 2007 farm densities in Puerto Rico (Y) were analyzed with conventional linear regression techniques. These densities are calculated by dividing the number of farms in by the area size of a municipality. Because the statistical distribution of these farm densities is highly skewed, they were subjected to a Box-Cox power transformation: $(Y - 0.12)^{0.38}$. The Shapiro-Wilk diagnostic test indicates that the transformed variable adequately mimics a normal distribution (p value = 0.5688). A benchmark bivariate linear regression model specification includes mean annual rainfall as a covariate variable. This allows a comparison between its parameter estimates and those for an ESF model.

[Table 75.5](#) summarizes estimation results for the bivariate and ESF linear regression models. The residual MC tests indicate that the linear regression results contain significant spatial autocorrelation (the MC z-score is 6.2299), whereas the ESF model successfully accounts for all but trace spatial autocorrelation (the MC z-score is -0.8873). A change in both regression coefficient (i.e., intercept and slope) estimates and their standard errors accompanies the change in model specification. The ESF model produces a larger estimate (0.0210 vis-à-vis 0.0138) for the covariate regression slope coefficient. Its significance within the context of the ESF model also is higher. Finally, the percent of variance accounted for (R^2) increases considerably with inclusion of the eight eigenvectors, which partly is due to an increase in the number of covariates: the mean annual rainfall variable alone explains 13%, whereas spatial autocorrelation as captured by the ESF term explains an additional 60%, of the variation in Y.

[Figure 75.5](#) portrays the decomposition of the 2007 farm densities based on the ESF model. [Figure 75.5a](#) is the map of the transformed farm densities. [Figure 75.5b](#) depicts the trend attributable to mean annual rainfall. This trend map illustrates that rainfall effectively explains the high values at the center of the island and in the eastern coast areas. [Figure 75.5c](#) portrays a spatially structured random component as captured by a linear combination of the eight selected eigenvectors. This map especially highlights the spatial clusters of low values in the northeast and of high values in the southwest parts of the island, indicating that they are well described by spatial autocorrelation map pattern components. [Figure 75.5d](#) visually confirms that the residuals of the ESF model lack a significant level of spatial autocorrelation.

Table 75.5 The results of linear regression for farm densities in Puerto Rico in 2007

Model features	Basic model		ESF model	
	Regression coefficient	Standard error	Regression coefficient	Standard error
Intercept	0.6622	0.3009*	0.1623	0.2235
Rainfall	0.0138	0.0042**	0.0210	0.0032***
# of selected eigenvectors	—		8	
R ²	0.1298		0.7283	
MC (z-score of MC)	0.4303 (6.2299)		-0.1644 (-0.8873)	

Significance codes: (*: 0.05, **: 0.01, ***: 0.001)

75.6 Extensions to Spatial Interaction Data Analysis

Spatial interaction can be conceptualized in a way that is analogous to gravity in Newtonian physics. The simple gravity model motivated its doubly constrained version containing origin and destination balancing factors (Wilson 1967). The balancing factors ensure that origin and destination totals are preserved. The factors may be interpreted as origin emissivity (i.e., a competitive accessibility measure vis-à-vis destinations with respect to origin i) and destination attractiveness (i.e., a competitive accessibility measure vis-à-vis origins with respect to destination j). This in turn motivated Poisson regression model estimation of its parameters (Flowerdew and Aitkin 1982). This Poisson probability model specification led to the use of origin and destination indicator variables [a separate indicator variable is included for each origin and each destination – that is, 2n 0–1 binary variables, each having n 1s and n(n–1) 0s], whose sets of coefficients are equivalent to the logarithm of the balancing factors when amalgamated. To avoid multicollinearity, coefficients for two of these indicator variables – one for the origin and one for the destination set – are set to 0, resulting in 2n–2 binary variables. Contemporary research again is focusing on the role spatial autocorrelation plays in this model (Chun 2008; Griffith 2011; Fischer and Griffith 2008; LeSage and Pace 2008): correlation exists among flows originating near origin areal unit i and terminating at destination areal unit j.

Each of the n^2 geographic flows tends to be positively correlated with its origin and destination sizes and negatively correlated with the extent of the intervening distance. The following simple equation furnishes a very good description of this phenomenon (see Griffith 2011):

$$F_{ij} \approx \kappa e^{SF_{O_i \times D_j}} A_i O_i B_j D_j e^{-\gamma d_{ij}} \quad (75.5)$$

where F_{ij} denotes the flow (e.g., number of workers) between locations i and j,
 κ is a constant of proportionality,

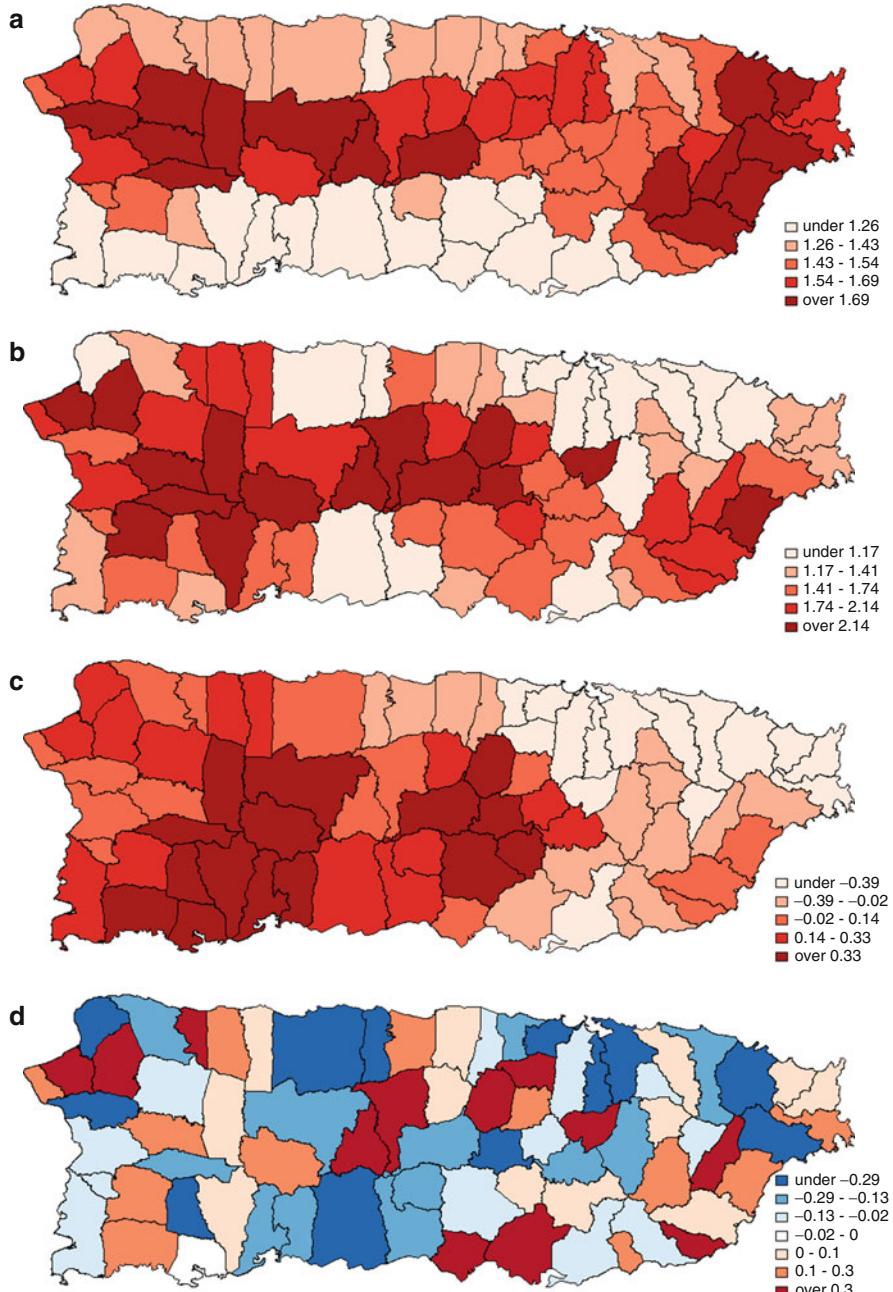


Fig. 75.5 Decomposed maps of 2007 farm densities in Puerto Rico: (a) dependent variable, (b) trend latent in the mean annual rainfall covariate, (c) spatially structured random component, and (d) random residuals

SF_{O_i} denotes the origin i spatial filter accounting for spatial autocorrelation in flows.

A_i denotes an origin balancing factor,

O_i denotes the total amount of flow leaving from origin i (e.g., number of workers residing in an origin),

SF_{D_j} denotes the destination j spatial filter accounting for spatial autocorrelation in flows,

B_j denotes a destination balancing factor,

D_j denotes the total amount of flow arriving to destination j (e.g., the number of jobs available in a destination),

d_{ij} denotes the distance separating origin i and destination j ,

γ denotes the global distance decay rate.

Selected results from the estimation of Eq. (75.5) for the Puerto Rican 2000 journey-to-work data (874,832 inter-municipality trips for 73² dyads) include the following:

Set values	$\hat{\kappa}$	$\hat{\gamma}$	Overdispersion	Pseudo-R ²
$SF_{O_i}=0, SF_{D_j}=0, A_i=1, B_j=1$	9.4×10^{-6}	0.1625	14.5227^2	0.8039
$SF_{O_i}=0, SF_{D_j}=0$	5.6×10^{-6}	0.2286	7.9801^2	0.9825
none	5.1×10^{-6}	0.2084	6.4750^2	0.9892

The spatial filter comprises 85 of 121 candidate eigenvectors (those with an MC of at least 0.25), from a total of 5,329 possible eigenvectors. The uncovered spatial autocorrelation in these flows contributes to excess Poisson variation. Adjusting for spatial autocorrelation in these flows yields a better alignment of the largest predicted and observed values (Fig. 75.6).

Figure 75.7 portrays the balancing factors and the spatial filters for the Puerto Rican journey-to-work example. Figures 75.7a and 75.7b, the A_i and B_j values, display conspicuous geographic patterns. Meanwhile, the origin spatial filter (Fig. 75.7c) contrasts the San Juan metropolitan region with the remainder of the island, whereas the destination spatial filter (Fig. 75.7d) highlights the four urban catchment areas (San Juan-Caguas, Arecibo, Mayaguez, and Ponce).

75.7 Conclusions

This chapter provides an introductory discussion about the concept of spatial autocorrelation together with ESF methodology. Spatial filtering furnishes a method to properly analyze a georeferenced variable by effectively separating a spatially structured random component from trend and random noise present in a georeferenced variable. Spatial filtering offers several advantages over other spatial data modeling methodologies, including spatial autoregression. First, it allows researchers to model spatially autocorrelated variables with conventional statistical tools, such as linear regression, which otherwise may well produce biased estimates with standard methodology, such as OLS. Second, it provides a way to

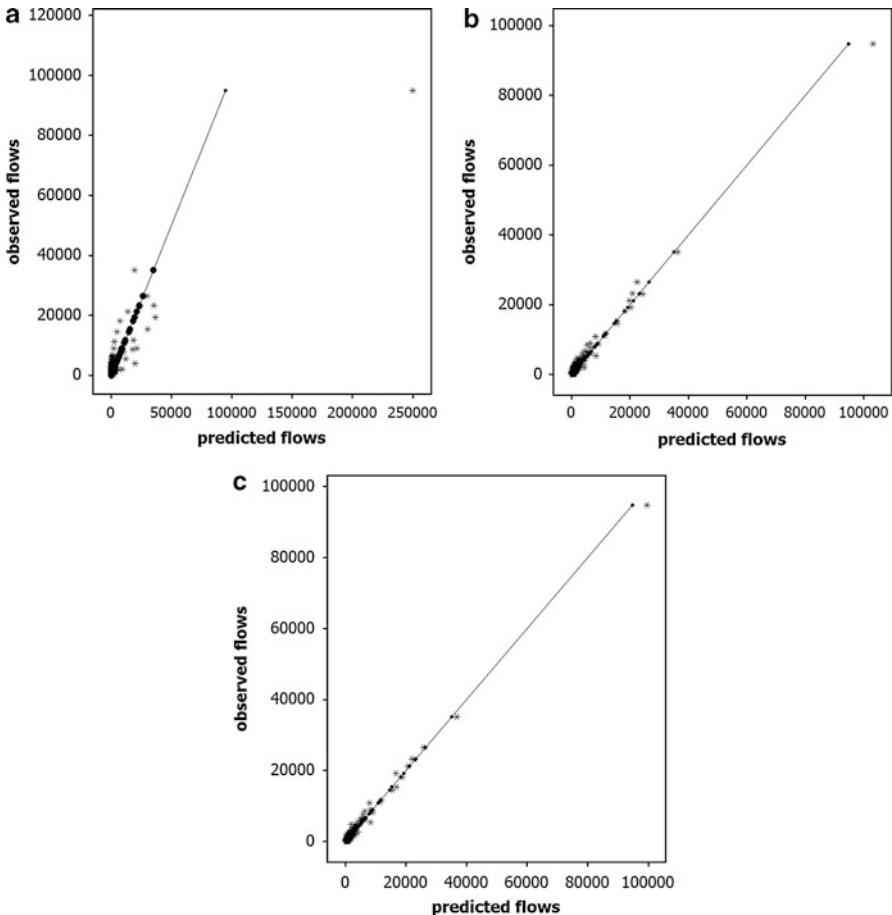


Fig. 75.6 Scatterplots of the Eq. (75.5)-predicted and observed journey-to-work trips. (a) $SF_{O_i} = 0$, $SF_{D_j} = 0$, $A_i = 1$, and $B_j = 1$. (b) $SF_{O_i} = 0$ and $SF_{D_j} = 0$. (c) all parameters estimated

analytically decompose a variable into underlying components, including a spatial component. Third, it provides a synthetic variate (the spatial filter) whose mapping visualizes spatial autocorrelation contained in a georeferenced variable. This visual representation can lead to a better understanding of geographical phenomenon, in part by furnishing a clue about possible variables missing from a linear regression model specification.

In addition, ESF methodology produces an enhanced and robust statistical result when compared with alternative model specifications, such as OLS linear regression and spatial autoregression: its parameters appear to be unbiased, relatively efficient, and consistent. Beyond the discussion in this chapter, the ESF methodology also offers the advantage of being able to accommodate positive spatial autocorrelation in a generalized linear model with discrete response variable,

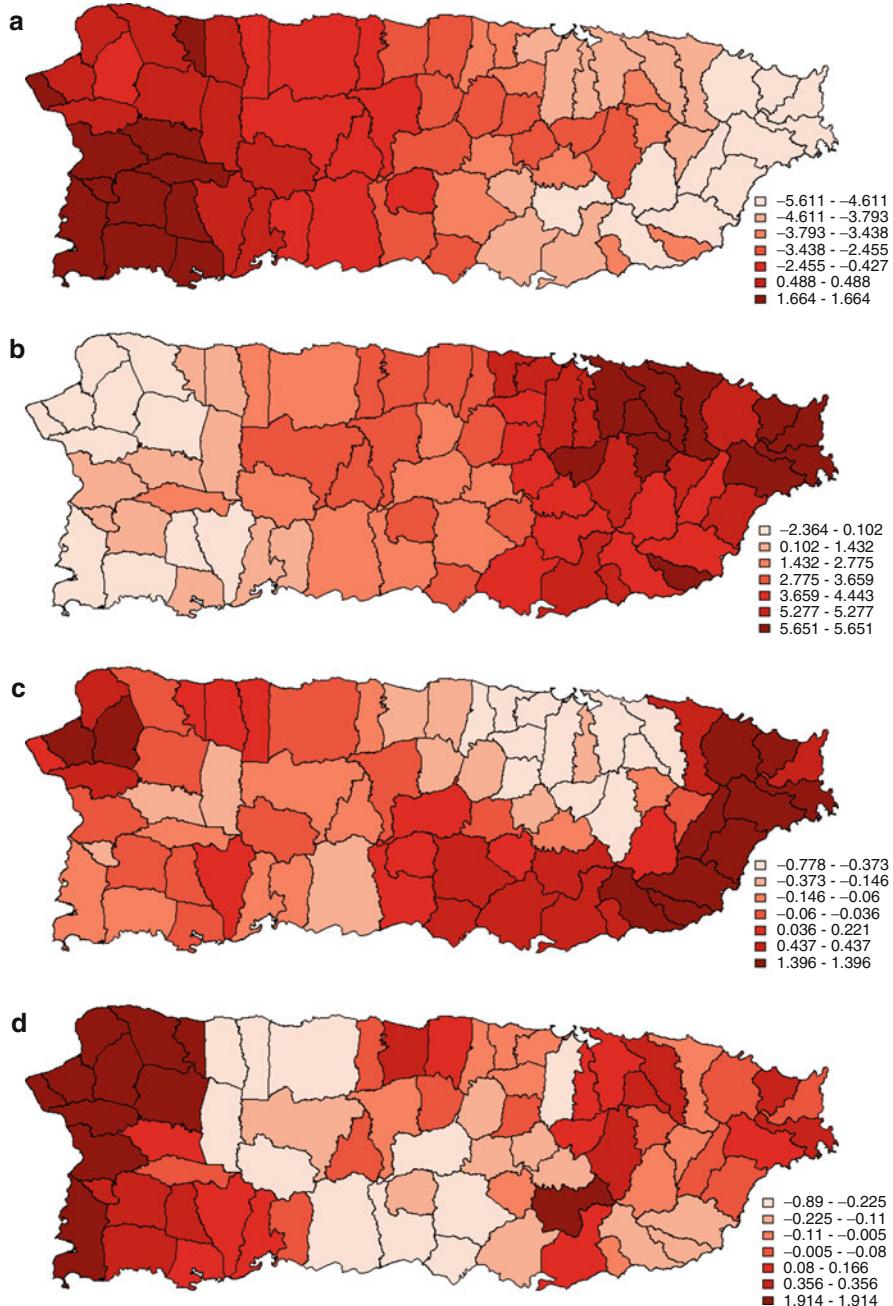


Fig. 75.7 Geographic distributions of Eq. (75.5) terms. (a) origin balancing factor (A_i). (b) destination balancing factor (B_j). (c) origin spatial filter. (d) destination spatial filter

such as Poisson and negative binomial, which cannot be modeled with the traditional auto specifications (Besag 1974). Further, the ESF methodology furnishes technology that can be utilized to account for spatial autocorrelation in geographical flows, such as population migration, journey-to-work flows, and interregional commodity flows. In other words, it offers a much wider array of tools, in more advanced spatial data analysis settings, than is addressed by the discussion in this chapter.

Finally, the ESF methodology can be extended to other spatial analysis problems. Spatial interaction modeling furnishes one such extension. Adjusting for spatial autocorrelation in geographic flow descriptions allows a better estimate of the global distance decay parameter, accounts for considerable excess variation characterizing flows, and better aligns the magnitudes of predicted and observed flows.

Appendix A

The relationship between the MC and a squared product moment correlation coefficient (r^2).

MC can be derived as a linear regression solution:

From OLS theory: $\mathbf{b} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}$

- (i) Convert the attribute variable in question to z-scores
- (ii) Let $\mathbf{X} = \mathbf{z}_Y$ and $\mathbf{Y} = \mathbf{Cz}_Y$
- (iii) Regress \mathbf{Cz}_Y on \mathbf{z}_Y , with a no-intercept option
- (iv) Let $\mathbf{X} = \mathbf{1}$ and $\mathbf{Y} = \mathbf{C1}$
- (v) Regress $\mathbf{C1}$ on $\mathbf{1}$, with a no-intercept option
- (vi) $MC = b_{\text{numerator}}/b_{\text{denominator}}$

This relationship relates directly to the Moran scatterplot, conveying why it is a useful visualization of spatial autocorrelation.

Next, let $MC = (n / \mathbf{1}^T \mathbf{C1}) \mathbf{z}^T \mathbf{Cz} / (n - 1)$ and rewrite vector \mathbf{z} as the following bivariate regression model specification: $\mathbf{z} = a\mathbf{1} + b\mathbf{CZ} + \mathbf{e}$, where \mathbf{e} is an n-by-1 vector of residuals. Then

$$\begin{aligned} b &= \frac{\mathbf{z}^T \mathbf{Cz}}{\mathbf{z}^T \mathbf{C}^2 \mathbf{z}} = \frac{s_z}{s_{Cz}} r = \frac{\sqrt{I}}{s_{Cz}} r \\ \frac{MC}{MC_1} \frac{n\lambda_1}{\mathbf{1}^T \mathbf{C1}} \frac{\mathbf{1}^T \mathbf{C1}}{n} \frac{n-1}{\mathbf{z}^T \mathbf{C}^2 \mathbf{z}} &= \frac{1}{\sqrt{\frac{\mathbf{z}^T \mathbf{C}(\mathbf{I} - \mathbf{1}\mathbf{1}^T/n)\mathbf{Cz}}{n-1}}} r \\ \left(\frac{MC}{MC_1} \right)^2 \lambda_1^2 \frac{(n-1)^2}{(\mathbf{z}^T \mathbf{C}^2 \mathbf{z})^2} &= \frac{n-1}{\mathbf{z}^T \mathbf{C}(\mathbf{I} - \mathbf{1}\mathbf{1}^T/n)\mathbf{Cz}} r^2 \\ r^2 &= \left(\frac{MC}{MC_1} \right)^2 \lambda_1^2 (n-1) \frac{\mathbf{z}^T \mathbf{C}(\mathbf{I} - \mathbf{1}\mathbf{1}^T/n)\mathbf{Cz}}{(\mathbf{z}^T \mathbf{C}^2 \mathbf{z})^2} \end{aligned}$$

where MC_1 denotes the maximum value of MC for a given spatial weight matrix C. For a large P-by-Q regular square lattice (i.e., $n = PQ$) and the rook's adjacency definition, for which $MC_1 \approx 1$, if $MC = 0.25$, then

$$\begin{aligned}\mathbf{1}^T \mathbf{C} \mathbf{z} &\approx 0 \\ \mathbf{z}^T \mathbf{C}^2 \mathbf{z} &\approx 16(PQ - 1)\end{aligned}$$

$$\lambda_1 = 2[\cos(\frac{\pi}{P+1}) + \cos(\frac{\pi}{Q+1})] \approx 4$$

and, consequently, $r^2 \approx 0.05$. Therefore, roughly 5% of the variance in a spatially autocorrelation random variable with $MC = 0.25$ is attributable to spatial autocorrelation.

References

- Besag J (1974) Spatial interaction and the statistical analysis of lattice systems. *J Roy Stat Soc Series B* 36(2):192–225
- Borcard D, Legendre P (2002) All-scale spatial analysis of ecological data by means of principal coordinates of neighbour matrices. *Ecol Model* 153:51–68
- Chun Y (2008) Modeling network autocorrelation within migration flows by eigenvector spatial filtering. *J Geogr Syst* 10(4):317–344
- Chun Y, Griffith DA (2009) Eigenvector selection with stepwise regression techniques to construct spatial filters. Paper presented at the annual association of american geographers meeting, Las Vegas, NV, 25 March
- Chun Y, Griffith DA (2011) Modeling network autocorrelation in space-time migration flow data: an eigenvector spatial filtering approach. *Ann Assoc Am Geogr* 101(3):523–536
- Cliff AD, Ord JK (1981) Spatial processes: models and applications. Pion, London
- Cordy C, Griffith DA (1993) Efficiency of least squares estimators in the presence of spatial autocorrelation. *Commun Stat Series B* 22:1161–1179
- de Jong P, Sprenger C, van Veen F (1984) On extreme values of Moran's I and Geary's c. *Geogr Anal* 16:17–24
- Fischer M, Griffith DA (2008) Modeling spatial autocorrelation in spatial interaction data: a comparison of spatial econometric and spatial filtering specifications. *J Reg Sci* 48:969–989
- Flowerdew R, Aitkin M (1982) A method of fitting the gravity model based on the Poisson distribution. *J Reg Sci* 22:191–202
- Getis A (1990) Screening for spatial dependence in regression analysis. *Pap Reg Sci Assoc* 69:69–81
- Getis A (2010) Spatial autocorrelation. In: Fischer M, Getis A (eds) *Handbook of applied spatial analysis*. Springer, New York, pp 255–278
- Getis A, Griffith DA (2002) Comparative spatial filtering in regression analysis. *Geogr Anal* 34(2):130–140
- Griffith DA (1992) What is spatial autocorrelation? Reflections on the past 25 years of spatial statistics. *l'Espace Géographique* 21:265–280
- Griffith DA (2000) Eigenfunction properties and approximations of selected incidence matrices employed in spatial analyses. *Linear Algebra Appl* 321:95–112
- Griffith DA (2003) *Spatial autocorrelation and spatial filtering: gaining understanding through theory and scientific visualization*. Springer, Berlin
- Griffith DA (2004) A spatial filtering specification for the autologistic model. *Environ Plann A* 36:1791–1811

- Griffith DA (2010) The Moran coefficient for non-normal data. *J Stat Plann Infer* 140:2980–2990
- Griffith DA (2011) Visualizing analytical spatial autocorrelation components latent in spatial interaction data: an eigenvector spatial filter approach. *Comput, Environ Urban Syst* 35:140–149
- Haining R (1991) Bivariate correlation with spatial data. *Geogr Anal* 23:210–227
- LeSage J, Pace R (2008) Spatial econometric modelling of origin–destination flows. *J Reg Sci* 48:941–967
- Pace K, LeSage J, Zhu S (2011) Interpretation and computation of estimates from regression models using spatial filtering. Paper presented to the Vth world conference of the spatial econometrics association, Toulouse, FR, 6–8 July
- Patuelli R, Griffith DA, Tiefelsdorf M, Nijkamp P (2011) Spatial filtering and eigenvector stability: space-time model for German unemployment data. *Int Reg Sci Rev* 34:235–280
- Portnoy S (1984) Asymptotic behavior of M estimators of p regression parameters when p^2/n is large: I. consistency. *Ann Stat* 12:1298–1309
- Tiefelsdorf M, Boots BN (1995) The exact distribution of Moran's I. *Environ Plann A* 27:985–999
- Tiefelsdorf M, Griffith DA (2007) Semi-parametric filtering of spatial autocorrelation: the eigenvector approach. *Environ Plann A* 39:1193–1221
- Tobler W (1969) A computer movie simulating urban growth in the Detroit region. Paper prepared for the meeting of the international geographical union, commission on quantitative methods, Ann Arbor, Michigan, August; published in 1970, *Economic Geography* 46(2) 234–240
- Tobler W (1975) Linear operators applied to areal data. In: Davis J, McCullagh M (eds) *Display and analysis of spatial data*. Wiley, New York, pp 14–37
- Wilson A (1967) A statistical theory of spatial distribution models. *Transport Res* 1:253–269

Section IX

Spatial Econometrics

spatial **matrix** **explanatory** **specification** **chain** **observations** **elements** **likelihood** **changes** **values** **parameters** **autoregressive** **spillovers** **Karthiga**

model **effects** **region** **inference** **test** **coefficient** **methods** **MCMC** **diagonal** **variance** **inference** **estimator** **dependent** **independent** **linear** **panel** **interaction** **regression** **dependence** **parameter** **estimator** **estimation** **indirect**

LeSage **variables** **flows** **distribution** **variable** **sample** **vector** **function** **distance** **lag** **scalar** **destination** **matrices** **dynamic** **autocorrelation** **fixed** **G** **neighbors** **Z** **regions** **Karthiga**

Julie Le Gallo

Contents

76.1	Introduction	1511
76.2	Spatial Effects in Cross-Sectional Models	1512
76.2.1	Forms of Spatial Autocorrelation in Regression Models	1513
76.2.2	Spatial Lag Model	1514
76.2.3	Cross-Regressive Model: Lagged Exogenous Variable	1516
76.2.4	Models with Spatial Error Autocorrelation	1517
76.2.5	Spatial Durbin Model	1520
76.2.6	Higher-Order Spatial Models	1520
76.2.7	Heteroscedasticity	1521
76.2.8	Parameter Instability	1522
76.3	Specification Tests in Spatial Cross-Sectional Models	1523
76.3.1	Moran's <i>I</i> Test	1524
76.3.2	Tests of a Single Assumption	1525
76.3.3	Tests in Presence of Spatial Autocorrelation or Spatial Lag	1526
76.3.4	Specification Search Strategies	1528
76.3.5	Non-nested Tests	1529
76.3.6	Spatial Autocorrelation and Spatial Heterogeneity	1529
76.4	Conclusion	1532
	References	1532

Abstract

This chapter provides a selective survey of specification issues in spatial econometrics. We first present the most commonly used spatial specifications in a cross-sectional setting in the form of linear regression models including a spatial lag and/or a spatial error term, heteroscedasticity or parameter instability. Second, we present a set of specification tests that allow checking

J. Le Gallo

CRESE, Université de Franche-Comté, Besançon, France

e-mail: jlegallo@univ-fcomte.fr

deviations from a standard, that is, nonspatial, regression model. An important space is devoted to unidirectional, multidirectional, and robust LM tests as they only require the estimation of the model under the null. Because of the complex links between spatial autocorrelation and spatial heterogeneity, we give some attention to the specifications incorporating both aspects and to the associated specification tests.

76.1 Introduction

In spatial regression models, the observations are collected from points or regions located in space. These models usually incorporate *spatial effects* that are commonly classified in two categories: *spatial autocorrelation* and *spatial heterogeneity*. On the one hand, spatial autocorrelation is a special case of cross-sectional dependence and refers to the coincidence of value similarity with locational similarity (Anselin and Bera 1998). Positive spatial autocorrelation means that observations from one location tend to exhibit values similar to those from nearby locations, while negative spatial autocorrelation points to the spatial clustering of dissimilar values. The typical characteristic of spatial autocorrelation is that it is two dimensional and multidirectional. On the other hand, spatial heterogeneity pertains to structural relations that vary over space, either in the form of nonconstant error variances in a regression model (heteroscedasticity) or in the form of spatially varying regression coefficients.

In recent years, the interest in *spatial econometrics*, that is, the subset of econometric methods that deals with the analysis of spatial effects in regression analysis, has seen an exponential growth in social sciences, leading to the creation of the *Spatial Econometrics Association* in 2006 (Arbia 2011). The upsurge in spatial econometrics has been driven by the recognition of the role of space and spatial/social interactions in economic theory, the availability of datasets with georeferenced observations, and the development of geographic information systems and spatial data analysis softwares. This field has even reached a stage of maturity through general acceptance as a mainstream methodology, according to Anselin (2010).

In this chapter, we provide a concise overview of the methodological issues related to the treatment of spatial effects in regression models. Attention here is given to specification issues, that is, how spatial correlation and spatial heterogeneity structures should be incorporated into a regression model and the implications for specification testing. We do not consider estimation issues, as this is the topic of other chapters in this volume (see Prucha and Jenish, ► Chap. 80, “Instrumental Variables/Method of Moments Estimation”; Mills and Parent, ► Chap. 79, “Bayesian MCMC Estimation” and Pace, ► Chap. 78, “Maximum Likelihood Estimation”). We have also limited the review to cross-sectional settings for linear regression models and do not consider spatial effects in space-time models (see Elhorst, ► Chap. 82, “Spatial Panel Models”) nor models for limited dependent variables (see Wang, ► Chap. 81, “Limited and Censored Dependent Variable Models”).

The chapter consists in two sections, starting with a presentation of the specification of spatial effects in cross-sectional linear regression models. Next, we consider specification tests that detect spatial autocorrelation and/or spatial heterogeneity. Most attention is devoted to spatial autocorrelation, the distinct nature of which requires a specialized set of techniques that are not a straightforward extension of time series methods to two dimensions. On the contrary, the treatment of spatial heterogeneity does not require specific econometric tools. However, we underline here the relationships between both effects. The chapter closes with some concluding remarks.

76.2 Spatial Effects in Cross-Sectional Models

Consider as a point of departure, the classical cross-sectional linear regression model:

$$y = X\beta + \varepsilon \quad (76.1)$$

where N is the total number of observations, here geographical areas; K is the total number of unknown parameters to estimate; y is the $(N,1)$ vector of observations on the dependent variable; X is the (N,K) matrix of observations on the K explanatory variables; β is the $(N,1)$ vector of unknown parameters to be estimated; and ε is the $(N,1)$ vector of error terms. We also assume that X is a non-stochastic matrix of full rank $K < N$.

If the error terms are $iid(0, \sigma^2 I_N)$, where I_N is the identity matrix of order N , then the Ordinary Least Squares (OLS) estimator defined by $\hat{\beta} = (X'X)^{-1}X'y$ is BLUE (Gauss-Markov theorem). However, the introduction of spatial effects in the linear regression model implies that some of these assumptions are not met. We first list the models incorporating some form of spatial autocorrelation and continue with models with spatial heterogeneity.

76.2.1 Forms of Spatial Autocorrelation in Regression Models

In the presence of spatial autocorrelation, the variance-covariance matrix in Eq. (76.1) $\Sigma = E(\varepsilon\varepsilon')$ contains N variances and $N(N - 1)/2$ off-diagonal parameters following a spatial ordering. These cannot be estimated separately with a cross section of N observations. Hence, in order to incorporate spatial autocorrelation in regression models, several possibilities exist. Some aim at imposing some structure or constraints on the elements of Σ such that a finite number of parameters characterizing spatial autocorrelation can be estimated. Others remain nonparametric. We briefly review these options here.

First, a *stochastic process* may be specified that determines the form of the covariance structure. In doing this, spatial lags are incorporated in the regression model. Spatial lags are obtained as the product of a *spatial weights matrix* W with

the vector of observations on a random variable. This matrix is of dimension (N,N) and specifies the connectivity structure within the observations in the sample. It has nonzero elements w_{ij} in each row i for those columns j that are neighbors of location j . The elements on the diagonal are equal to 0. The notion of *neighbors* can be purely geographic, such as sharing a common border, or can be more general, such as neighbors in social network space. Spatial autocorrelation is then modeled by specifying various functional relationships between the vector of observations of the explained variable y and its spatial lag Wy , a spatially lagged error term We and/or spatially lagged explanatory variables WX .

Second, the covariance between observations can be specified as a direct and continuous function of distance. Different specifications have been suggested.

Third, a nonparametric approach can be adopted where the functional form of the function of distance separating two equations is left unspecified. This can also accommodate heteroscedasticity of unknown form.

We detail these different possibilities below.

76.2.2 Spatial Lag Model

In this model, labeled SAR model, spatial autocorrelation is incorporated through a *spatial lag* of the *endogenous* variable. The *structural* model is written as

$$\begin{aligned} y &= \rho Wy + X\beta + \varepsilon \\ \varepsilon &\rightarrow iid(0, \sigma^2 I_N) \end{aligned} \tag{76.2}$$

Wy is the endogenous lag variable for the spatial weights matrix W ; ρ is the spatial autoregressive parameter that indicates the strength of interactions existing between the observations of y .

In the spatial lag model, observation y_i is, in part, explained by the values taken by y in neighboring observations: $(Wy)_i = \sum_{j \neq i} w_{ij}y_j$. Indeed, when W is standardized, each element $(Wy)_i$ is interpreted as a weighted average of the y values for i 's neighbors. The introduction of Wy allows evaluating the degree of spatial dependence when the impact of other variables is controlled for. When Eq. (76.2) is the result of a theoretical modeling implying some process of social and spatial interaction, this parameter measures substantive spatial dependence, that is, the extent of spatial externalities or spatial diffusion.

Symmetrically, it allows controlling spatial dependence when evaluating the impact of other explanatory variables. In this case, particular care should be given to the interpretation of the coefficient estimates (see below).

LeSage and Pace (2009) provide several motivations for regression models that include a spatial lag. One is a time-dependence motivation: cross-sectional model relations with a spatial lag may come from economic agents considering past period behavior of neighboring agents. The presence of a spatial lag has also been justified with theoretical models involving diffusion, copycatting,

or spatial externalities. These are the cases of substantive spatial dependence. It is then the formal representation of the equilibrium outcome of spatial interaction processes.

Note that ρ is not a conventional correlation coefficient between vector y and its spatial lag Wy . Indeed, this parameter is not restricted to the range -1 to 1 . From the DGP associated with the SAR model, the log-likelihood function involves a Jacobian term of the form $\ln|I_N - \rho W|$ that constrains the parameter ρ to be in the interval $[1/w_{\min}; 1/w_{\max}]$ where w_{\min} and w_{\max} are respectively the minimum and the maximum eigenvalues of W . If the latter is row standardized, then $w_{\max} = 1$.

When a spatial lag variable is ignored in the model specification, whereas it is present in the underlying data generating process, the OLS estimators in the spatial model Eq. (76.1) are biased and not consistent (omitted variable bias).

This specification has several properties:

76.2.2.1 Multiplier and Diffusion Effects

Assume that the matrix $(I_N - \rho W)$ is not singular. In this case, Eq. (76.2) can be rewritten in the following *reduced* form:

$$y = (I_N - \rho W)^{-1}X\beta + (I_N - \rho W)^{-1}\varepsilon \quad (76.3)$$

This model is nonlinear in ρ and β . It follows from Eq. (76.3) that $E(y) = (I_N - \rho W)^{-1}X\beta$. The matrix inverse $(I_N - \rho W)^{-1}$ is a full matrix and not triangular, as in the time series case where dependence is only one directional. When $|\rho| < 1$, this implies an infinite series, the *Leontief expansion*, involving the explanatory variables and the error term at all locations:

$$y = (I_N + \rho W + \rho^2 W^2 + \dots)X\beta + (I_N + \rho W + \rho^2 W^2 + \dots)\varepsilon \quad (76.4)$$

This expression allows defining two effects: a *multiplier* effect affecting the explanatory variables and a *spatial diffusion* effect affecting the error terms.

On the one hand, with respect to the explanatory variables, this expression means that in average, the value of y at one location i is not only explained by the values of the explanatory variables associated to this location but also by those associated to all the other locations (neighbors or not) *via* the inverse spatial transformation $(I_N - \rho W)^{-1}$. This spatial multiplier effect decreases with distance, that is, the powers of W in the series expansion of $(I_N - \rho W)^{-1}$.

On the other hand, with respect to the error process, this expression means that a random shock in a location i not only affects the value of y in this location but also has an impact on the values of y in all the other locations *via* the same spatial inverse transformation. This is the diffusion effect, which also declines with distance.

Both these effects are *global* in the sense that all locations in the system interact with each other (Anselin 2003).

From Eq. (76.3), it also follows that $E[(Wy)_i \varepsilon_i] = E\left[\left\{W(I_N - \rho W)^{-1} \varepsilon\right\}_i \varepsilon_i\right] \neq 0$. The spatial lag is therefore always endogenous, irrespective of the properties of ε , so that the estimation of model Eq. (76.2) cannot be based on OLS but should be performed using maximum likelihood (ML), instrumental variables (IV), or Bayesian methods.

76.2.2.2 Interpretation of Coefficient Estimates

A consequence of the multiplier effect in the spatial lag model is that particular care should be taken when interpreting the coefficient estimates (LeSage and Pace, ► Chap. 77, “Interpreting Spatial Econometric Models” for more details). Indeed, the impact of a marginal change in one variable X_k on $E(y)$ is not equivalent to the coefficient associated to X_k , noted β_k , as in the standard regression model. On the contrary, it follows from Eq. (76.3) that

$$\frac{\partial E(y_i)}{\partial X_{jk}} = S_k(W)_{ij} \quad (76.5)$$

where X_{jk} is the value of X_k at location j and $S_k(W)_{ij}$ is the ij^{th} element of the matrix $(I_N - \rho W)^{-1} \beta_k$. Hence, the impact of a change in an explanatory variable differs over all observations. Summary measures of these impacts are discussed in LeSage and Pace (2012).

76.2.2.3 Variance-Covariance Matrix

From Eq. (76.3), we derive the variance-covariance matrix of y :

$$E(yy') = (I_N - \rho W)^{-1} E(\varepsilon\varepsilon') (I_N - \rho W')^{-1} \quad (76.6a)$$

$$E(yy') = \sigma^2 (I_N - \rho W)^{-1} (I_N - \rho W')^{-1} \quad (76.6b)$$

This variance-covariance matrix is full, which implies that each location is correlated with every other location in the system. However, this correlation decreases with distance.

76.2.2.4 Endogenous Spatial Lag and Heteroscedasticity

Note $u = (I_N - \rho W)^{-1} \varepsilon$. Its variance-covariance is written as

$$E(uu') = \sigma^2 (I_N - \rho W)^{-1} (I_N - \rho W')^{-1} \quad (76.7)$$

Equation (76.7) shows that the covariance between each pair of error terms is not null and decreasing with the order of proximity. Moreover, the elements of the diagonal of $E(uu')$ are not constant. This implies error *heteroscedasticity* of u , whether or not ε is heteroscedastic.

76.2.3 Cross-Regressive Model: Lagged Exogenous Variable

Another possibility to incorporate spatial autocorrelation in a regression model is to include one or more *exogenous lagged variables* in Eq. (76.1):

$$\begin{aligned} y &= \rho W y + W Z \delta + \varepsilon \\ \varepsilon &\rightarrow iid(0, \sigma^2 I_N) \end{aligned} \quad (76.8)$$

Z is a matrix of dimension (N,L) containing L variables that may or not correspond to the variables included in X ; WZ is the matrix of observations for the exogenous lagged variables with weights matrix W , and δ is the $(L,1)$ vector of spatial parameters indicating the intensity of spatial correlation existing between the observations in y and those of Z .

In this model, the observation y_i is explained by the values taken by the variables in X in location i and by the variables in Z in neighboring regions. The interactions in the system hence remain *local*.

Contrary to the spatial lag model and the models with a spatial error autocorrelation (below), the estimation of the cross-regressive model can be based on OLS.

76.2.4 Models with Spatial Error Autocorrelation

Finally, spatial autocorrelation can be incorporated in a regression model by specifying a spatial process in the error terms. It is therefore a special form of a nonspherical error variance-covariance matrix with $E[\varepsilon_i \varepsilon_j] \neq 0$ for two locations $i \neq j$. As such, these models should be estimated using ML, generalized method of moments (GMM), or Bayesian methods. The different possibilities lead to different error spatial covariances that differ with respect to the range and extent of spatial interaction in the model.

76.2.4.1 Spatial Autoregressive Process

The most commonly used specification is a spatial autoregressive process in the error terms. The structural model can be then written as

$$\begin{aligned} y &= X\beta + \varepsilon \\ \varepsilon &= \lambda W \varepsilon + u \end{aligned} \quad (76.9)$$

The parameter λ is the spatial autoregressive coefficient that reflects the interdependence between the regression residuals; u is the error term such as $u \rightarrow iid(0, \sigma^2 I_N)$. When spatial error autocorrelation is omitted, the OLS estimators are unbiased, but inefficient estimators and the statistical inference based on OLS are biased.

This specification has several properties:

Spatial Diffusion

First, if the matrix $(I_N - \lambda W)$ is not singular, then model Eq. (76.9) can be rewritten under the following reduced form:

$$y = X\beta + (I_N - \lambda W)^{-1}u \quad (76.10)$$

This expression leads to a global spatial diffusion effect as in model Eq. (76.3) but, as $E(y) = X\beta$, there is no spatial multiplier effect.

Variance-Covariance Matrix

From Eq. (76.10), we have

$$E(yy') = E(\varepsilon\varepsilon') = (I_N - \rho W)^{-1}E(\varepsilon\varepsilon')(I_N - \rho W')^{-1} \quad (76.11a)$$

$$E(yy') = E(\varepsilon\varepsilon') = \sigma^2(I_N - \rho W)^{-1}(I_N - \rho W')^{-1} \quad (76.11b)$$

Hence, we find, for ε and for y , a structure identical to that of the spatial lag model: this process leads to nonzero error covariance between each pair of observations, but these covariances decrease with distance. The spatial structure of the variance-covariance induced by the model with spatial error autocorrelation is therefore *global*, since it links all the locations of the system to all others.

Moreover, the error structure induces nonconstant elements of the diagonal of $E(\varepsilon\varepsilon')$, which implies heteroscedasticity of the errors ε , whether u is heteroscedastic or not.

Constrained Spatial Durbin Model

Model Eq. (76.9) can be rewritten in a form where both an endogenous spatial lag and all exogenous spatial lags appear. Indeed, by multiplying both sides of Eq. (76.10) by $(I_N - \lambda W)$ and moving the autoregressive term to the right, we obtain the constrained *spatial Durbin model*:

$$y = \lambda Wy + X\beta - \lambda WX\beta + u \quad (76.12)$$

This specification shows how the spatial error model is a special case of a spatial lag model, with additional nonlinear constraints on the parameters. This forms the basis of a specification test that will be presented below.

Several alternatives have been suggested in the literature even if their application is less frequent in the literature.

76.2.4.2 Spatial Moving-Average Process

The spatial *moving-average* process is specified as

$$\begin{aligned} y &= X\beta + \varepsilon \\ \varepsilon &= \gamma Wu + u \end{aligned} \quad (76.13)$$

where γ is the moving-average coefficient and u is the error term such as $u \rightarrow iid(0, \sigma^2 I_N)$. Contrary to the previous case, the reduced model does not contain any inverse matrices since Eq. (76.13) already corresponds to the reduced model. The variance-covariance matrix resulting from this process is

$$E(\varepsilon\varepsilon') = \sigma^2(I_N + \gamma W)(I_N + \gamma W') = \sigma^2[I_N + \gamma(W + W') + \gamma^2 WW'] \quad (76.14)$$

In contrast to the variance-covariance matrix associated with the autoregressive process, Eq. (76.14) is not a full matrix. The nonzero covariances only exist for first-order ($W + W'$) and second-order (WW') neighbors. This process therefore implies much less overall interaction than the autoregressive model, and the spatial structure of covariance induced by Eq. (76.14) is only *local* since it does link all the locations of system to each other.

Finally, as in the autoregressive case, the elements of the diagonal of Eq. (76.14) are not constant, implying, as in the previous model, heteroscedasticity in ε , irrespective of the nature of u .

76.2.4.3 Kelejian and Robinson Specification

Kelejian and Robinson (1995) suggest another specification in which the error term is the sum of two independent terms, one being a smoothing term of neighboring errors and the other being specific to the location:

$$\varepsilon = Wu + v \quad (76.15)$$

where u and v are supposed homoscedastic and independent. Then, the variance-covariance matrix of ε is

$$E(\varepsilon\varepsilon') = \sigma_v^2 I_N + \sigma_u^2 WW' = \sigma^2[I_N + \varphi WW'] \quad (76.16)$$

where σ_u^2 and σ_v^2 are the variance, respectively, associated with u and v , $\sigma^2 = \sigma_v^2 > 0$ and $\varphi = \sigma_u^2/\sigma_v^2$. The spatial interaction implied by Eq. (76.16) is more limited than in the moving-average model as it only concerns neighbors of the first and second order contained in the nonzero elements of WW' . Heteroscedasticity is also implied in this specification.

76.2.4.4 Direct Representation and Nonparametric Specifications

In this case, the covariance between each pair of error terms is directly specified as an inverse function of the distance between them: $\text{cov}(\varepsilon_i, \varepsilon_j) = \sigma^2 f(\theta, d_{ij})$ where d_{ij} is the distance between i and j , σ^2 is the error variance, and f is the distance function. This function is a distance decay function that should ensure definite-positive variance-covariance matrix. This imposes constraints on the functional form, the parameter space, the metric, and scale used for the distance measure. For instance, one might use a negative exponential distance decay function:

$$E(\varepsilon\varepsilon') = \sigma^2[I_N + \gamma\Psi] \quad (76.17)$$

where the off-diagonal elements of Ψ are given by $\Psi = e^{-\theta d_{ij}}$ where θ is a nonnegative scaling parameter. The diagonal elements of Ψ are set to zero.

Contrary to the previous specifications, the direct representation does not induce heteroscedasticity.

An alternative to parametric specifications is to leave the functional form unspecified: these are *nonparametric* models. We then have $\text{cov}(\varepsilon_i, \varepsilon_j) = f(d_{ij})$ where d_{ij} is a positive and symmetric distance metric. The regularity conditions on the distance metric have been derived by Conley (1999).

The presence of spatial error autocorrelation is often interpreted as a problem in the model specification, such as functional form problems or spatial autocorrelation resulting from a mismatch between the spatial scale of the phenomenon being studied and the spatial scale at which it is measured.

76.2.5 Spatial Durbin Model

An encompassing specification to the spatial lag model, the spatial cross-regressive model, and the spatial error model is the unconstrained spatial Durbin model. The latter contains a spatially lagged endogenous variable and all the spatially lagged exogenous variables. More specifically, it is written as

$$y = \lambda W y + X\beta + \lambda W X \delta + u \quad (76.18)$$

The spatial lag model, the spatial cross-regressive model, and the spatial error model are found with the appropriate constraints on the parameters, respectively, $H_0 : \delta = 0$, $H_0 : \rho = \delta = 0$, and $H_0 : \lambda\beta + \delta = 0$.

LeSage and Pace (2009) provide several motivations for a spatial Durbin model. One is an *omitted variable* motivation. Indeed, they show that if the linear regression model Eq. (76.1) is affected by an omitted variables problem and if these omitted variables are spatially correlated and correlated with the included explanatory variables, then unbiased estimates of the coefficients associated with the endogenous variables X can still be obtained by fitting a spatial Durbin model. Other motivations detailed in LeSage and Pace (2009) are based on spatial heterogeneity and model uncertainty.

76.2.6 Higher-Order Spatial Models

In these models, multiple spatially lagged dependent variables and/or multiple spatially lagged error terms are included.

For instance, the spatial autoregressive, moving-average SARMA(p,q) process is as follows:

$$\begin{aligned} y &= X\beta + \rho_1 W_1 y + \rho_2 W_2 y + \dots + \rho_p W_p y + \varepsilon \\ \varepsilon &= \lambda_1 W_1 u + \lambda_2 W_2 u + \dots + \lambda_p W_p u + u \end{aligned} \quad (76.19)$$

In general, the weights W_i are associated to the i^{th} order of contiguity. We could similarly consider a process where the errors follow a spatial autoregressive process of order q . However, in this case, identification issues may arise (Anselin 1988).

It may be that these *high-order processes* are the result of a poorly specified spatial weights matrix rather than a realistic data generating process (Anselin and Bera 1998). For instance, if the weights matrix of the model underestimates the real spatial interaction in the data, there will be residual spatial error autocorrelation. This can lead to the estimation of higher-order processes while only a well-specified weights matrix should be necessary. These higher-order models are in fact usually used as alternatives in diagnostic tests. Rejection of the null may then indicate that a different specification of the weights is necessary.

76.2.7 Heteroscedasticity

Until now, all specifications have assumed *iid* innovations. However, as we have seen, the sole presence of spatial autocorrelation induces *heteroscedasticity* in the models. In cross-sectional regression, additional heteroscedasticity is also frequently present. For instance, in the spatial autoregressive error model, we can have

$$\begin{aligned} y &= X\beta + \varepsilon \\ \varepsilon &= \lambda W\varepsilon + u \\ u &\rightarrow iii(0, \Omega) \end{aligned} \tag{76.20}$$

In this case, the variance-covariance matrix of ε is

$$E(\varepsilon\varepsilon') = (I_N - \rho W)^{-1} \Omega (I_N - \rho W')^{-1} \tag{76.21}$$

Several specifications have been used for Ω . In a spatial context, a useful one is that of *groupwise heteroscedasticity*. When the data are organized into spatial regimes, one variance is estimated for each regime so that Ω has a block-diagonal structure:

$$\Omega = \begin{bmatrix} \sigma_1^2 I_{N_1} & 0 & \cdots & 0 \\ 0 & \sigma_2^2 I_{N_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_2^2 I_{N_2} \end{bmatrix} \tag{76.22}$$

where L is the number of regimes, N_l , $l = 1 \dots L$ is the number of observations in regime l , and I_{N_l} , $l = 1 \dots L$ is the identity matrix of dimension N_l .

The variance can also be specified as a function of variables:

$$\sigma_i^2 = \sigma^2 f(z'_i \alpha) \tag{76.23}$$

where σ^2 is a scale parameter, f is some functional form, and z_i is a $(P, 1)$ vector of variables and α_i , $i = 1 \dots P$ are unknown parameters to estimate. For instance, in a spatial context, Casetti and Can (1999) suggest the DARP (Drift Analysis of Regression Parameters) model: the variance of the error terms is expanded into a monotonic function of the observations' distance from a reference point in an expansion space:

$$\sigma_i^2 = e^{\gamma_0 + \gamma_1 h_i} \quad (76.24)$$

where h_i is the square of the distance between the i^{th} observation and one reference point (such as the Central Business District in a city).

The variance-covariance matrix can also be left unspecified as in the nonparametric approach. For instance, Kelejian and Prucha (2007) suggest a nonparametric heteroscedasticity- and autocorrelation-consistent (*HAC*) estimator of the variance-covariance matrix in a spatial context, that is, a SHAC procedure. They assume that the $(N, 1)$ disturbance vectors ε of model Eq. (76.1) are generated as follows: $\varepsilon = R\xi$ where R is a (N, N) non-stochastic matrix whose elements are not known. This disturbance process allows for general patterns of correlation and heteroscedasticity. The asymptotic distribution of the corresponding OLS or instrumental variable (IV) estimators implies the variance-covariance matrix $\psi = N^{-1}Z'\Sigma Z$, where $\Sigma = (\sigma_{ij})$ denotes the variance-covariance matrix of ε . Kelejian and Prucha (2007) show that the SHAC estimator for its $(r, s)^{\text{th}}$ element is

$$\hat{\psi}_{rs} = N^{-1} \sum_{i=1}^N \sum_{j=1}^N x_{ir}x_{js}\hat{\varepsilon}_i\hat{\varepsilon}_j K(d_{ij}/d_n) \quad (76.25)$$

where x_{ir} is the i^{th} element of the r^{th} explanatory variable, $\hat{\varepsilon}_i$ is the i^{th} element of the OLS or IV residual vector, d_{ij} is the distance between unit i and unit j , d_n is the bandwidth, and $K(\cdot)$ is the kernel function with the usual properties.

76.2.8 Parameter Instability

Spatial heterogeneity can also manifest by *parameter instability*, that is, the lack of constancy in some, or all, of the parameters in the regression model. This instability has a spatial dimension: the regression coefficients correspond to a number of distinct spatial regimes. The spatial variability of the coefficients can be discrete, if systematic differences between regimes are observed. In this case, model coefficients are allowed to vary between regimes. It can also be continuous over space.

In the absence of spatial autocorrelation, the case of discrete spatial heterogeneity can be readily treated with standard tools such as dummy variables, ANOVA, or spline functions. Recently, some authors have investigated the possibility of spatial heterogeneity affecting the spatial lag or spatial error coefficients. In this

case, the methodology consists in estimating higher-order models where the spatial matrices pertain to different spatial regimes rather than different order of contiguities.

Heterogeneity can also be continuous. In this case, rather than partitioning the cross-sectional sample into regimes, we assume that parameter heterogeneity is location specific. One possibility is to use *geographically weighted regression*, labeled GWR (Fotheringham et al. 2004), which is a locally linear, nonparametric estimation method. The base model for one location i is

$$y_i = \sum_{k=1}^K \beta_{ki} x_{ki} + \varepsilon_i \quad (76.26)$$

A different set of parameters is estimated for each observation by using the values of the characteristics taken by neighboring observations. With respect to spatial autocorrelation, Pace and LeSage (2004) have pointed out that if spatial autocorrelation only arises due to inadequately modeled spatial heterogeneity, GWR can potentially eliminate the problem. However, this is not necessarily the case when substantive interactions coexist with parameter heterogeneity. Therefore, Pace and LeSage (2004) have generalized GWR to allow simultaneously for spatial parameter heterogeneity and spatial autocorrelation: the *spatial autoregressive local estimation* (SALE):

$$U(i)y = \rho_i U(i)Wy + U(i)X\beta_i + U(i)\varepsilon \quad (76.27)$$

where $U(i)$ represents a (N,N) diagonal matrix containing distance-based weights for observation i that assigns the weights of one to the m nearest neighbors to observation i and weights of zero to all the other observations. The product $U(i)y$ then represents a $(m,1)$ subsample of observations on the explained variables associated with the m observations nearest in location to observation i . The other products are interpreted in a similar fashion. As $m \rightarrow N$, $U(i) \rightarrow I_N$, the local estimates approach the global estimates from the SAR model as the subsample increases.

76.3 Specification Tests in Spatial Cross-Sectional Models

Ignoring spatial effects when it is present have various effects on the estimates' properties. It may lead to biased and inconsistent estimates of the model parameters for an omitted spatial lag or inefficient estimated and biased inference for omitted spatial error autocorrelation and/or omitted heteroscedasticity. Hence, specification testing is therefore relevant in applied work and constitutes the topic of this section.

We first present Moran's I test, where the alternative is an unspecified form of spatial autocorrelation. Second, we detail the most commonly used tests of spatial autocorrelation based on *maximum likelihood*: tests of a single alternative,

conditional tests, and robust tests. Indeed, as featured in ► Chap. 80, “Instrumental Variables/Method of Moments Estimation” and ► Chap. 78, “Maximum Likelihood Estimation”, there might be some complexities involved in the estimation of spatial processes, based on nonlinear optimization (maximum likelihood or generalized methods of moments). Consequently, tests based on the *Lagrange multiplier* (LM) principle (or score test) have been extensively used in specification testing. Contrary to *Wald* (W) or *likelihood ratio* (LR) tests, they only necessitate the estimation of the model under the null hypothesis, typically the simple regression model as in Eq. (76.1). We also briefly present tests based on alternative principles. Third, some strategies aimed at finding the best specification have been devised, when the researcher does not have an a priori of the form taken by spatial autocorrelation. Finally, we outline the complex interactions between spatial autocorrelation and spatial heterogeneity and present how spatial heterogeneity can be tested.

76.3.1 Moran's *I* Test

Moran's *I* test is a *diffuse test* as the alternative is not a specified form of spatial autocorrelation. It is the two-dimensional analog of the test of temporal correlation in univariate time series for regression residuals (Moran 1950). In matrix notations, it is formally written as

$$I = \frac{N}{S_0} \left(\frac{e' We}{e'e} \right) \quad (76.28)$$

where $e = y - X\tilde{\beta}$ is the vector of OLS regression residuals, W is the spatial weights matrix, and S_0 is a standardization factor equal to the sum of all elements of W . For a row-standardized weights matrix W , this element simplifies to 1. The first two moments under the null were derived by Cliff and Ord (1972):

$$E(I) = \frac{tr(MW)}{N - K} \quad (76.29)$$

$$V(I) = \frac{tr(MWMW') + tr(MW)^2 + \{tr(MW)^2\}}{(N - K)(N - K + 2)} - [E(I)]^2 \quad (76.30)$$

where M is the usual symmetric and idempotent matrix : $M = I_N - X(X'X)^{-1}X'$. Inference is then based on the standardized value: $Z(I) = [I - E(I)]/V(I)$. For normally distributed residuals, $Z(I)$ asymptotically follows a centered normal distribution. Under the null assumption of spatial independence, Moran's *I* test is locally best invariant and is also asymptotically equivalent to a likelihood ratio of $H_0 : \lambda = 0$ in Eq. (76.9) or of $H_0 : \gamma = 0$ in Eq. (76.13); it therefore shares the asymptotic properties of these statistics. Moreover, Moran's *I* has power against any alternative of spatial correlation, including a spatial lag alternative.

In the remainder of the section, we consider tests with a specific alternative, that is, focused tests, and concentrate on Lagrange multiplier tests that only require the estimation of the model under the null hypothesis. Some of these tests are *unidirectional* when the alternative deals with one specific misspecification; others are *multidirectional* when the alternative comprises various misspecifications.

76.3.2 Tests of a Single Assumption

76.3.2.1 Spatial Error Autocorrelation

First, consider the case where the error terms follow a spatial autoregressive model Eq. (76.9): $\varepsilon = \lambda W\varepsilon + u$. We test $H_0 : \lambda = 0$. The null corresponds to the linear classical model Eq. (76.1). The multiplier Lagrange statistic can be written the following way (Anselin 1988):

$$LM_{ERR} = \frac{[e'We/(e'e/N)]^2}{T} \quad (76.31)$$

where $T = \text{tr}[(W' + W)W]$, tr is the trace operator, and e is the vector of OLS regression residuals. This is equivalent to a scaled Moran coefficient. Since there is only one constraint, under the null, this statistic is asymptotically distributed as a $\chi^2(1)$.

The test statistic is the same if we specify as alternative assumption the moving-average process Eq. (76.13) with the test $H_0 : \gamma = 0$. LM_{ERR} is therefore locally optimal for the two alternatives (autoregressive and moving average). Consequently, when the null is rejected, the test does not provide any indications with respect to the form of the error process.

Pace and LeSage (2008) argue that the test of spatial error autocorrelation can be performed using a Hausman test, since under the null (model 1), there are two consistent estimators differing in efficiency (OLS and ML), and under the alternative (model 2) only one estimator is efficient (ML).

76.3.2.2 Kelejian-Robinson Specification

For the specification of the error suggested by Kelejian and Robinson (1995), a Lagrange multiplier test can also be derived following the same principle. Using notations of model Eq. (76.15), testing the null $H_0 : \varphi = 0$ yields a statistic of the form (Anselin 2001)

$$KR = \left[\frac{e'W'We}{e'e/N} - T_1 \right]^2 \Big/ 2 \left[T_2 - \frac{T_1^2}{N} \right] \quad (76.32)$$

where $T_1 = \text{tr}(WW^2)$ and $T_2 = \text{tr}(WW'WW')$. Under the null, this statistic is asymptotically distributed as a $\chi^2(1)$.

76.3.2.3 Common Factor Test

The *common factor test* allows choosing between a model with spatial error autocorrelation and a spatial Durbin model. The unconstrained spatial Durbin model in Eq. (76.18) and the spatial error model in Eq. (76.9) are equivalent if $H_0 : \lambda\beta + \delta = 0$. This test can be performed with the Lagrange multiplier principle. The corresponding statistic is asymptotically distributed as a $\chi^2(K - 1)$.

76.3.2.4 Test of an Endogenous Spatial Lag

In this case, the null hypothesis is $H_0 : \rho = 0$ in Eq. (76.2). The test statistic is (Anselin 1988)

$$LM_{LAG} = \frac{[e'Wy/(e'e/N)]^2}{D} \quad (76.33)$$

with $D = (WX\tilde{\beta})'M(WX\tilde{\beta})/\tilde{\sigma}^2 + \text{tr}(W'W + WW)$ where $\tilde{\beta}$ and $\tilde{\sigma}^2$ are the OLS estimates. This statistic is asymptotically distributed as a $\chi^2(1)$.

76.3.3 Tests in Presence of Spatial Autocorrelation or Spatial Lag

In specification testing, it is useful to know if the model contains both a spatial error autocorrelation and an endogenous spatial lag. In this respect, Anselin et al. (1996) note that LM_{ERR} is the test statistic corresponding to $H_0 : \lambda = 0$ when assuming a correct specification for the rest of the model, that is, $\rho = 0$. However, if $\rho \neq 0$, this test is not valid anymore, even asymptotically as it is not distributed as a centered χ^2 . Hence, valid statistical inference necessitates taking account of a possible endogenous variable when testing spatial error autocorrelation and vice versa.

Facing this problem, three strategies are possible. First, one can perform a *joint test* of the presence of an endogenous spatial lag and a spatial error autocorrelation. However, if the null is rejected, the exact nature of spatial dependence is not known. Second, another solution consists in estimating a model with an endogenous spatial lag and then tests for residual spatial autocorrelation and vice versa. Third, Anselin et al. (1996) suggest *robust tests* based on OLS residuals in the simple model but that are capable of taking account a spatial error autocorrelation when testing endogenous spatial lag and vice versa.

76.3.3.1 Joint Test

The first approach is the test of the joint null hypothesis $H_0 : \lambda = \rho = 0$ in a model containing both a spatial lag and a spatial error:

$$\begin{aligned} y &= \rho W_1 y + X\beta + \varepsilon \\ \varepsilon &= \lambda W_2 \varepsilon + u \end{aligned} \quad (76.34)$$

The Lagrange multiplier test is based on the OLS residuals. The test statistic is (Anselin 1988)

$$SARMA = \frac{\left[(\tilde{d}_\lambda)^2 D + (\tilde{d}_\rho)^2 T_{22} - 2\tilde{d}_\lambda \tilde{d}_\rho T_{12} \right]}{DT_{22} - T_{12}^2} \quad (76.35a)$$

or

$$SARMA = \frac{\tilde{d}_\lambda}{T} + \frac{(\tilde{d}_\lambda - \tilde{d}_\rho)^2}{D - T} \text{ if } W_1 = W_2 \quad (76.35b)$$

where $\tilde{d}_\lambda = (e'We)/(e'e/n)$, $\tilde{d}_\rho = (e'Wy)/(e'e/n)$, and $T_{ij} = \text{tr}[W_i W_j + W'_j W_j]$. Under the null, SARMA is asymptotically distributed as a $\chi^2(2)$. If the null is rejected, the exact nature of spatial dependence is not known. Extensions of these principles to joint tests in SARMA (p,q) models are derived in Anselin (2001).

76.3.3.2 Conditional Tests

This approach consists in performing a Lagrange multiplier test for a form of spatial dependence when the other form is not constrained. For instance, we test $H_0 : \lambda = 0$ in presence of ρ . The null corresponds to the spatial lag model, whereas the alternative corresponds to Eq. (76.31). The test is then based on the residuals of model Eq. (76.2) estimated by maximum likelihood. The test statistic is as follows (Anselin 1988):

$$LM_{ERR}^* = \frac{\hat{d}_\rho^2}{T_{22} - (T_{21A})^2 \hat{V}(\hat{\rho})} \quad (76.36)$$

where $T_{21A} = \text{tr}[W_2 W_1 A^{-1} + W'_2 W_1 A^{-1}]$, $A = I_N - \hat{\rho} W_1$, $\hat{\rho}$ is the maximum likelihood estimator of ρ , and $\hat{V}(\hat{\rho})$ is the estimated variance of $\hat{\rho}$ in model Eq. (76.2). Under the null, this statistic is asymptotically distributed as a $\chi^2(1)$.

Conversely, we can also $H_0 : \rho = 0$ in presence of λ ; the test is then based on the maximum likelihood $\hat{\theta}$ in the spatial error model Eq. (76.9). The statistic is (Anselin 1988)

$$LM_{LAG}^* = \frac{(\hat{e}B'BW_1y)^2}{H_\rho - H_{\theta\rho}\hat{V}(\hat{\theta})H'_{\theta\rho}} \quad (76.37)$$

where $\theta = (\beta', \lambda, \sigma^2)$, $\hat{\theta}$ is the maximum likelihood estimator of θ , $B = I_N - \hat{\lambda}W_1$, and $\hat{V}(\hat{\theta})$ is the estimated variance-covariance matrix of $\hat{\theta}$ in model Eq. (76.9). The other terms are

$$H_\rho = \text{tr}(W_1^2) + \text{tr}(BW_1B^{-1}) + \frac{(BW_1X\hat{\beta})' (BW_1X\hat{\beta})}{\hat{\sigma}^2} \quad (76.38)$$

$$H_{\theta\rho} = \text{tr} \begin{bmatrix} \frac{(BX)'BW_1X\hat{\beta}}{\hat{\sigma}^2} \\ \text{tr}(W_2B^{-1})BW_1B^{-1} + \text{tr}(W_2W_1B^{-1}) \\ 0 \end{bmatrix} \quad (76.39)$$

Under the null, this statistic is asymptotically distributed as a $\chi^2(1)$.

76.3.3.3 Robust Tests

The third approach, suggested by Anselin et al. (1996), consists in using robust tests to a local misspecification. For instance, LM_{ERR} is adjusted so that its asymptotic distribution remains a centered $\chi^2(1)$, even in local presence of ρ . This test can be done using the OLS residuals of the simple model Eq. (76.1). Assuming $W_1 = W_2$, the modified statistic for the test $H_0 : \lambda = 0$ is

$$RLM_{ERR} = \frac{(\tilde{d}_\lambda - TD^{-1}\tilde{d}_\rho)^2}{[T(1 - TD)]} \quad (76.40)$$

Similarly, the test statistic of $H_0 : \rho = 0$ in local presence of λ is

$$RLM_{LAG} = \frac{(\tilde{d}_\lambda - \tilde{d}_\rho)^2}{D - T} \quad (76.41)$$

76.3.4 Specification Search Strategies

Tests based on Lagrange multiplier have been very popular in applied spatial econometrics in *specification search*, as they only require the estimation of the model under the null, typically, the simple model estimated by OLS. They can be combined to develop a *specific-to-general* sequential specification search strategy, that is, a forward stepwise specification search, whenever no a priori spatial specification has been chosen.

The first step consists in estimating the simple model Eq. (76.1) by means of OLS and in performing Moran's I test and the SARMA test. The rejection of the null in both cases indicates omitted spatial autocorrelation but not the form taken by this autocorrelation.

If the null hypothesis is rejected, it may be a sign of model misspecification. For instance, using a Monte Carlo experiment, McMillen (2003) shows that incorrect functional forms or omitted variables that are correlated over space might produce spurious spatial autocorrelation. It may therefore be useful to include in the model, if possible, additional variables. It can be exogenous additional variables that may eliminate or reduce spatial dependence, or exogenous spatial lags, corresponding in total or in part to the initial explanatory variables.

If the addition of exogenous variables has not eliminated spatial autocorrelation, a model incorporating a spatial lag and/or a spatial error must be estimated.

The choice between these two forms of spatial dependence can be done by comparing the significance levels of LM_{ERR} Eq. (76.31) and LM_{LAG} Eq. (76.33) and their robust versions RLM_{ERR} Eq. (76.40) and RLM_{LAG} Eq. (76.41): if LM_{LAG} (resp. LM_{ERR}) is more significant than LM_{ERR} (resp. LM_{LAG}) and RLM_{LAG} (resp. RLM_{ERR}) is significant but not RLM_{ERR} (resp. RLM_{LAG}), a spatial lag (resp. a spatial error) must be included in the regression model (Anselin and Florax 1995).

Once the spatial lag or the spatial error model has been estimated, three additional tests can be implemented. On the one hand, for a spatial lag model, LM_{ERR}^* allows checking whether an additional spatial error is still necessary. On the other hand, for a spatial error model, LM_{LAG}^* allows checking whether an additional spatial lag is still necessary. The common factor test allows checking whether the restriction $H_0 : \lambda\beta + \delta = 0$ is rejected or not. If not, Eq. (76.18) reduces to the spatial error model Eq. (76.9).

There are several drawbacks with this classical specific-to-general approach. First, the significance levels of the sequence of tests are unknown. Second, every test is conditional on arbitrary assumptions that may be tested later. The inference is then invalid if these assumptions are indeed rejected. As a consequence, the results of this approach is subject to the order in which the tests are carried out and whether or not adjustments are made in the significance levels of the sequence of tests.

Alternatively, a *general-to-specific* search strategy, that is, a forward stepwise specification search, can be implemented based on the spatial Durbin model Eq. (76.18) as it encompasses most spatial specifications. Model Eq. (76.18) is estimated, and testing is performed using Wald statistics or likelihood ratio statistics. Then, the failure to reject the common factor constraints suggests a spatial error model, while rejection of these constraints suggests a spatial lag model. In the first case, the significance of the spatial error coefficient is tested; if it is significant, the final specification is the error model Eq. (76.9); if it is not, the final model is the simple model Eq. (76.1). Likewise, in the second case, the significance of the spatial lag coefficient is tested; if it is not significant, the final model selection is the standard regression model. Simulation experiments performed by Florax et al. (2003) compare the specific-to-general and the general-to-specific strategies and provide some evidence of better performances of the forward strategy, in terms of power and accuracy.

76.3.5 Non-nested Tests

The basis of these specification search strategies above is that the competing models are nested within a more general model (spatial Durbin model). However, for non-nested alternatives, other strategies must be devised. For instance, Kelejian and Piras (2011) have extended the J -test procedure to a spatial framework. The null hypothesis corresponds to a spatial error-spatial lag model as in Eq. (76.34) with similar weights, while the alternative hypothesis corresponds to a set of G models that differ with the model in H_0 with respect to the specification of the regressor matrix, the weighting matrix, the disturbance term, or a combination of these three.

76.3.6 Spatial Autocorrelation and Spatial Heterogeneity

Spatial autocorrelation and spatial heterogeneity are often both present in regressions. We have already underlined that heteroscedasticity is implied by the presence of a spatial lag or a spatial error term. More generally, these two effects entertain complex links. First, there may be observational equivalence between these two effects in a cross section (Anselin and Bera 1998). Secondly, heteroscedasticity and structural instability tests are not reliable in the presence of spatial autocorrelation. Conversely, spatial autocorrelation tests are affected by heteroskedasticity. Thirdly, spatial autocorrelation is sometimes the result of unmodeled parameter instability. In other words, if space-varying relationships are modeled within a global regression, the error terms may be spatially autocorrelated. All these elements suggest that both aspects cannot be considered separately. We briefly review here some tests that have tackled this issue.

76.3.6.1 Spatial Autocorrelation and Heteroscedasticity

First, a joint test of spatial error autocorrelation and heteroscedasticity consists in the sum of a Breusch-Pagan test and the LM_{ERR} (Anselin 1988). The resulting statistic is asymptotically distributed as a $\chi^2(P)$, where P is the number of variables that affect the variance (Eq. 76.23). Alternatively, Kelejian and Robinson (1998) derive a joint test for spatial autocorrelation and heteroscedasticity that does not require the normality assumption for the error terms and the regression model to be linear.

Conditional tests may also be performed. On the one hand, a Lagrange multiplier test of spatial autocorrelation in a regression with heteroscedastic error terms may be derived. Let $\hat{\Omega}$ be the estimated diagonal variance-covariance matrix, then the heteroscedastic LM statistics becomes (Anselin 1988):

$$LM = \frac{(e' \hat{\Omega}^{-1} W e)^2}{tr(WW + W' \hat{\Omega}^{-1} W \hat{\Omega})} \quad (76.42)$$

where e is the vector of residuals in the heteroscedastic regression. This statistic is asymptotically distributed as a $\chi^2(1)$.

On the other hand, a test of heteroscedasticity in a spatial lag model or a spatial error model can be performed. In the first case, a Breusch-Pagan statistic is computed on the ML residuals, while in the second case, it is performed on spatially filtered residuals in the ML estimation.

76.3.6.2 Spatial Autocorrelation and Parameter Instability

In the case of discrete parameter heterogeneity under the form of spatial regimes in a homoscedastic model, a test of equality of some or all parameters between regimes can be performed using a standard Chow test. However, when error spatial

autocorrelation and/or heteroscedastic is present, this must be adjusted. Formally, without loss of generality, consider a model with two regimes:

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} X_1 & 0 \\ 0 & X_2 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \end{bmatrix} \quad (76.43)$$

Let $\varepsilon = [\varepsilon'_1 \ \varepsilon'_2]$ and the variance-covariance matrix: $\Psi = E(\varepsilon\varepsilon')$. The test of parameter stability is $H_0 : \beta_1 = \beta_2$.

When $\Psi = \sigma^2 \Omega$, then the test statistic is (Anselin 1988)

$$C_G = \frac{\hat{e}'_c \hat{\Omega}^{-1} \hat{e}_c - \hat{e}'_L \hat{\Omega}^{-1} \hat{e}_L}{\hat{\sigma}^2} \quad (76.44)$$

where \hat{e}_c is the vector of estimated residuals of the constrained model and \hat{e}_L the vector of estimated residuals of the unconstrained residuals. This statistic is asymptotically distributed as a $\chi^2(K)$, where K is the number of explanatory variables in the model.

Whenever the break affects the spatial coefficient, Mur et al. (2010) suggest LM tests. For instance, assume a spatial lag model where a simple break (such a center vs. periphery) only affects the parameter of spatial dependence:

$$\begin{aligned} y &= \rho_0 W y + \rho_1 W^* y + X\beta + \varepsilon \\ \varepsilon &\rightarrow iid(0, \sigma^2 I_N) \end{aligned} \quad (76.45)$$

where ρ_0 is the spatial lag coefficient pertaining to the second regime, ρ_1 represents the difference between the first regime and the second regime, and W^* is a weights matrix defined as $w_{ij}^* = w_{ij}$ if location i or location j belongs to the first regime and $w_{ij}^* = 0$ otherwise. Then the LM statistic for the test $H_0 : \rho_1 = 0$ is

$$LM_{LAG}^{BREAK} = \frac{\left[\frac{y' W^* \tilde{\varepsilon}}{\hat{\sigma}^2} - \text{tr} \tilde{A}^{-1} W^* \right]^2}{\hat{\sigma}^2} \quad (76.46)$$

where $\tilde{\varepsilon}$ is the vector of residuals of the ML estimation of Eq. (76.2), $\tilde{\sigma}^2$ is the corresponding estimated variance, $\tilde{A} = I_N - \tilde{\rho}W$ where $\tilde{\rho}$ is the ML estimation in Eq. (76.2), and $\hat{\sigma}^2$ is the ML estimated variance corresponding to the linear restriction of the null. This statistic is asymptotically distributed as a $\chi^2(1)$.

A spatial error model with a structural break affecting the spatial error parameter is

$$\begin{aligned} y &= X\beta + \varepsilon \\ \varepsilon &= \lambda_0 W \varepsilon + \lambda_1 W^* \varepsilon + u \\ u &\rightarrow iid(0, \sigma^2 I_N) \end{aligned} \quad (76.47)$$

The LM statistic for the test $H_0 : \lambda_1 = 0$ is as follows:

$$LM_{LAG}^{BREAK} = \frac{\left[\frac{\tilde{e}' W^* \tilde{B} \tilde{e}}{\hat{\sigma}^2} - \text{tr} \tilde{B}^{-1} W^* \right]^2}{\hat{\sigma}^2} \quad (76.48)$$

where \tilde{e} is the vector of residuals of the ML estimation of Eq. (76.9), $\tilde{\sigma}^2$ is the corresponding estimated variance, $\tilde{B} = I_N - \tilde{\lambda}W$ where $\tilde{\lambda}$ is the ML estimation in Eq. (76.9), $\hat{\sigma}^2$ is the ML estimated variance corresponding to the linear restriction of the null. This statistic is asymptotically distributed as a $\chi^2(1)$.

76.4 Conclusion

The objective of this chapter was to provide a concise review of specification issues in spatial econometrics. We focused on the way spatial effects may be incorporated into regression models and on specification testing. We first presented the most commonly used spatial specifications in a cross-sectional setting in the form of linear regression models including a spatial lag and/or a spatial error term, heteroscedasticity, or parameter instability. Second, we presented a set of specification tests that allow checking deviations from a standard, that is, nonspatial, regression model. An important space has been devoted to LM tests as they only require the estimation of the model under the null. Unidirectional, multidirectional, and robust LM tests are now in the standard toolbox of spatial econometrics. They are still frequently used in applied work, even though the technical/numerical difficulties associated to the estimation of spatial models have become much more tractable, even for very large samples. Because of the complex links between spatial autocorrelation and spatial heterogeneity, we have given some attention to the specifications incorporating both aspects and to the associated specification tests.

References

- Anselin L (1988) Spatial econometrics, methods and models. Kluwer, Dordrecht
- Anselin L (2001) Rao's score test in spatial econometrics. J Stat Plan Infer 97:113–139
- Anselin L (2003) Spatial externalities, spatial multipliers and spatial econometrics. Int Reg Sci Rev 26:153–166
- Anselin L (2010) Thirty years of spatial econometrics. Pap Reg Sci 89:3–25
- Anselin L, Bera AK (1998) Spatial dependence in linear regression models with an application to spatial econometrics. In: Ullah A, Giles DEA (eds) Handbook of applied economics statistics. Springer, Berlin
- Anselin L, Florax RGJM (1995) Small sample properties of tests for spatial dependence in regression models: some further results. In: Anselin L, Florax RJGM (eds) New directions in spatial econometrics. Springer, Berlin
- Anselin L, Bera A, Florax RGJM, Yoon M (1996) Simple diagnostic test for spatial dependence. Reg Sci Urban Econ 26:77–104

- Arbia G (2011) A lustrum of SEA: recent research trends flowing the creation of the spatial econometrics association (2007–2011). *Spat Econ Anal* 6:377–395
- Casetti E, Can A (1999) The econometric estimation and testing of DARP models. *J Geogr Syst* 1:91–106
- Cliff A, Ord JK (1972) Testing for spatial autocorrelation among regression residuals. *Geogr Anal* 4:267–284
- Conley TG (1999) GMM estimation with cross-sectional dependence. *J Econom* 92:1–44
- Florax RJGM, Folmer H, Rey SJ (2003) Specification searches in spatial econometrics: the relevance of Hendry's methodology. *Reg Sci Urban Econ* 33:557–579
- Fotheringham AS, Brundson C, Charlton M (2004) Geographically weighted regression: the analysis of spatially varying relationships. Wiley, Chichester
- Kelejian HH, Piras G (2011) An extension of Kelejian's J-test for non-nested spatial models. *Reg Sci Urban Econ* 41:281–292
- Kelejian HH, Prucha I (2007) HAC estimation in a spatial framework. *J Econom* 140:131–154
- Kelejian HH, Robinson DP (1995) Spatial correlation: a suggested alternative to the autoregressive model. In: Anselin L, Florax RJGM (eds) *Advances in spatial econometrics*. Springer, Heidelberg
- Kelejian HH, Robinson DP (1998) A suggested test for spatial autocorrelation and/or heteroskedasticity and corresponding Monte-Carlo results. *Reg Sci Urban Econ* 28:389–417
- LeSage J, Pace KP (2009) *Introduction to spatial econometrics*. CRC Press, Boca Raton
- McMillen DP (2003) Spatial autocorrelation or model misspecification? *Int Reg Sci Rev* 26:208–217
- Moran P (1950) A test for the serial dependence of the residuals. *Biometrika* 35:255–260
- Mur J, Lopez F, Angulo A (2010) Instability in spatial error models: an application to the hypothesis of convergence in the European case. *J Geogr Syst* 12:259–280
- Pace RK, LeSage JP (2008) A spatial Hausman test. *Econ Lett* 101:282–284
- Pace RK, LeSage JP (2004) Spatial autoregressive local estimation. In: Getis A, Mur J, Zoller H (eds) *Spatial statistics and spatial econometrics*. Palgrave MacMillan

James P. LeSage and R. Kelley Pace

Contents

77.1	Introduction	1536
77.2	Spatial Regression Models	1537
77.2.1	Spatial Error Models	1537
77.2.2	Spatial Lag of X Models	1539
77.2.3	Spatial Lag of y Models	1541
77.2.4	Measures of Dispersion for the Effects Estimates	1544
77.2.5	Partitioning Global Effects Estimates Over Space	1545
77.3	Applications of Spatial Regression Models	1546
77.3.1	Spatial Error Models	1546
77.3.2	SLX and SDEM Models	1548
77.3.3	SAR and SDM Models	1549
77.4	Conclusion	1551
	References	1551

Abstract

Past applications of spatial regression models have frequently interpreted the parameter estimates of models that include spatial lags of the dependent variable incorrectly. A discussion of issues surrounding proper interpretation of the estimates from a variety of spatial regression models is undertaken. We rely on scalar summary measures proposed by LeSage and Pace (Introduction to

J.P. LeSage (✉)

Department of Finance and Economics, Texas State University – San Marcos,
San Marcos, TX, USA

e-mail: jlesage@spatial-econometrics.com

R.K. Pace

Department of Finance, E.J. Ourso College of Business Administration, Louisiana State
University, Baton Rouge, LA, USA
e-mail: kelley@pace.am

spatial econometrics. Taylor Francis/CRC Press, Boca Raton, 2009) who motivate that these reflect a proper interpretation of the *marginal effects* for the nonlinear models involving spatial lags of the dependent variable. These nonlinear spatial models are contrasted with linear spatial models, where interpretation is more straightforward. One of the major advantages of spatial regression models is their ability to quantify spatial spillovers. These can be defined as situations where nonzero cross-partial derivatives exist that reflect impacts on outcomes in region i arising from changes in characteristics of region j . Of course, these cross-partial derivatives can be interpreted as impacts of changes in an own region characteristic on other regions or changes in another regions' characteristic on the own region. The ability to produce empirical estimates along with measures of dispersion that can be used for inference regarding the statistical significance, magnitude, and spatial extent of spillovers provides a major motivation for using spatial regression models.

77.1 Introduction

Spatial spillovers reflect a major theme in regional science. A loose definition of spillovers in a spatial context would be that changes occurring in one region exert impacts on other regions. For example, changes in tax rates by one jurisdiction might exert an impact on tax rate setting decisions of nearby regions, a phenomenon that has been labeled tax mimicking and yardstick competition between local governments (Allers and Elhorst 2005; Deskins and Hill 2010). Other examples include situations where home improvements made by one homeowner exert a beneficial impact on the selling prices of neighboring homes, innovation by university researchers diffuses to nearby firms, air or water pollution generated in one region spills over to nearby regions, etc. We will provide a more formal definition of spatial spillovers in this chapter.

It would be of interest to be able to test for the presence of statistically significant spatial spillovers and to quantify the magnitude and spatial extent of these if they exist. For example, in the context of tax mimicking, a test for the statistical significance of spillovers where changes in tax rates in region i exert a statistically significant effect on future tax rate changes in other regions' $j \neq i$ would provide evidence regarding the presence or absence of tax mimicking. Knowing the sign and magnitudes of spillovers would aid in discerning the nature of the mimicking behavior. An empirical estimate of the spatial extent of the spillovers would also be useful for studying this type of phenomena. For example, how many neighboring regions (on average over the sample of regions observed) are impacted as a result of a change in tax rates by the typical region i ? Spatial regression models provide one way to obtain answers to questions of this type.

We draw a distinction between *global and local spillovers*, an idea discussed in Anselin (2003). Assuming that possible connections exist among regions, global spillovers arise when changes in a characteristic of one region impact all regions' outcomes. This applies even to the region itself since impacts can pass to the

neighbors and back to the own region (feedback). Specifically, global spillovers impact the neighbors, neighbors to the neighbors, neighbors to the neighbors to the neighbors, and so on. Local spillovers represent a situation where the impacts fall only on nearby or immediate neighbors, dying out before they impact regions that are neighbors to the neighbors. Therefore, the *feedback effects* that arise in the case of global spillovers do not accompany local spillovers.

Feedback effects arise when changes to own region/entity characteristics exert an impact on outcomes in the own and neighboring regions/entities, which produce additional changes or feedback effects on outcomes in the own region. As an example, when a homeowner A improves the value of their property, this exerts a beneficial impact on the selling price of home A plus that of neighboring homes, say B . However, an increase in the selling price of neighboring homes B will produce a beneficial feedback effect on the selling price of home A . These feedback effects have sometimes been labeled *self-reinforcing effects* or virtuous cycles, with regional economic growth often being characterized as this type of phenomena. Growth can start with an exogenous technological innovation that leads to learning curve effects and economies of scale. This in turn leads to reduced costs and improved production efficiencies which result in lower average market prices. As prices decrease, consumption and aggregate output increase, and with increased levels of output, there are more learning and scale effects that start a new cycle. The feedback effects of global spillovers and the absence of these for local spillovers make it useful to draw a distinction between alternative approaches that can be used to model these two phenomena.

77.2 Spatial Regression Models

We discuss interpretation of the *marginal effects estimates* and how they relate to spatial spillovers for a host of spatial regression models in this section of this chapter. Applied studies in the literature that use the various models for the purpose of drawing inferences about local and global spillovers are discussed in the next section of the chapter.

77.2.1 Spatial Error Models

The *spatial error model* (SEM) (Ord 1975; Anselin 1988) and *spatial moving average* (SMA) (Haining 1990; Fingleton 2001) error models are shown in Eqs. (77.1) and (77.2), where we introduce these models as a contrast to other models that will be discussed. These models do not allow for spatial spillovers arising from changes in characteristics of one region on outcomes observed in other regions. We are relying on a definition of spillovers introduced by LeSage and Pace (2009) who define spatial spillovers as nonzero cross-partial derivatives $\partial y_j / \partial x_i$. This means that changes to explanatory variables in region i impact the dependent variable values in region $j \neq i$:

$$y = X\beta + u, \quad u = (I_n - \rho W)^{-1}\varepsilon \quad (77.1)$$

$$y = X\beta + u, \quad u = (I_n + \theta W)\varepsilon \quad (77.2)$$

In these equations, the $n \times 1$ vector y represents a cross-sectional dependent variable that exhibits variation across spatial observational units, and the $n \times k$ matrix X represents explanatory variables that usually include a vector of ones. The scalar parameters ρ and θ measure the strength of spatial dependence with boundaries on the permissible (stationary) parameter space determined by minimum and maximum eigenvalues of the $n \times n$ matrix W (see LeGallo, ▶ Chap. 76, “Cross-Section Spatial Regression Models,” for details concerning the permissible parameter space). For simplicity, we assume that W has all real eigenvalues and that the principal eigenvalue equals 1. The matrix W provides a (normalized) structure of connectivity between the observations, and in spatial regression models, *each observation is a region*. In a spatial context, connectivity might be defined as *neighboring regions* using nonzero elements in the i,j th position of the matrix W to denote that region j is a neighbor to region i . The matrix W has row sums of one and a main diagonal with zeros (so regions cannot be neighbors to themselves). The $n \times 1$ vector ε is a disturbance term usually assumed to be normally distributed with zero mean, constant variance σ^2 , and zero covariance across observations.

The parameters of the models are β , ρ , θ , and σ^2 which can be estimated using maximum likelihood, Bayesian, or instrumental variable methods (see ▶ Chap. 78, “Maximum Likelihood Estimation,” Mills and Parent ▶ Chap. 79, “Bayesian MCMC Estimation” and Prucha and Jennish ▶ Chap. 80, “Instrumental Variables/Method of Moments Estimation” for details concerning estimation).

For both the SEM and SMA models, the cross-partial derivatives in Eq. (77.4) (spillovers) are zero by design, as in the case of nonspatial regression models. The SEM model estimate $\hat{\beta}_r$ for the r th variable in the explanatory variables matrix X (and associated measure of dispersion) forms the basis for inference regarding how changes this explanatory variable in region i will impact the i th region values of the dependent variable, and this scalar estimate averages over all $i = 1, \dots, n$ observations. As noted, for nonlinear models, we need to rely on marginal effects when interpreting parameter estimates, rather than the coefficient estimates associated with parameters β of the model. In the case of the SEM model, the parameter estimate equals the average marginal effect of the own variable (which LeSage and Pace (2009) label the average direct effect) on the dependent variable y . Further, the average marginal effect of the spillovers (which LeSage and Pace (2009) label the average indirect effect) is 0, as shown in Eq. (77.4):

$$\partial y_i / \partial x_i^r = \beta_r \quad (77.3)$$

$$\partial y_j / \partial x_i^r = 0 \quad (77.4)$$

These models do allow for diffusion of shocks or disturbances across observations/regions. This can be seen by considering the matrix inverse expression for the SEM disturbances, which can be expressed using an infinite series expansion: $(I_n - \rho W)^{-1} = (I_n + \rho W + \rho^2 W^2 + \rho^3 W^3 + \dots) \varepsilon$. The matrix inverse $(I_n - \rho W)^{-1}$ exists under the typical assumptions made regarding the scalar parameter ρ that measures the strength of spatial dependence in the dependent variable and the spatial weight matrices W employed in these models (see ► Chap. 76, “Cross-Section Spatial Regression Models”).

If we consider a scalar shock δ to a single region i , reflected by $\varepsilon_i + \delta$, which results in a new vector $\tilde{\varepsilon}$, then we have $(I_n + \rho W + \rho^2 W^2 + \rho^3 W^3 + \dots) \tilde{\varepsilon}$ as the new disturbance vector. The first term $I_n \tilde{\varepsilon}$ will exert an impact on the disturbance for region i , whereas the second term $\rho W \tilde{\varepsilon}$ will impact disturbances of regions that neighbor observation i . This is because the matrix–vector product $W \tilde{\varepsilon}$ produces a resulting vector that reflects a linear combination of shocks from observations neighboring each observation. The change in disturbance $\varepsilon_i + \delta$ of observation i will be included in the linear combination of shocks from observations that neighbor i . Powers of the matrix W when used to form matrix–vector products such as $W^2 \tilde{\varepsilon}$ will form linear combinations based on neighbors to the neighbors of each observation, so the shock to observation i will exert an impact on *second-order* neighboring observations to i . Second-order neighbors are neighbors to the neighbors. A similar statement applies to higher-order powers $W^3 \tilde{\varepsilon}, W^4 \tilde{\varepsilon}$, and so on; these form linear combinations involving neighbors to the neighbors, neighbors to the neighbors to the neighbors, etc. The implication is that the SEM model allows for diffusion of shocks that arise for a single observation to other observations, with a decay of influence for higher-order neighbors. The decrease in magnitude of impact for higher-order neighbors is a consequence of the fact that $\rho < 1$ and the principal eigenvalue of the row-normalized matrix W is one. The SEM model allows for *global diffusion* of shocks to other observations. We note also that feedback effects arise here, consistent with the distinction made earlier about global versus local phenomena. This can be seen by recognizing that the matrix W contains zeros on the main diagonal, but the matrix W^2 does not. Since W^2 contains nonzero elements in positions reflecting neighbors to neighbors of each observation, and each observation is a neighbor to its neighbor, there are nonzero elements on the diagonal. These nonzero diagonal elements capture the feedback effect of the diffusion arising from a shock to region i . Of course, similar statements apply to higher powers of the matrix W .

The SMA model allows for *local diffusion* of a shock to the disturbance of region i . Using our vector $\tilde{\varepsilon}$, we have $(I_n + \theta W) \tilde{\varepsilon} = \tilde{\varepsilon} + \theta W \tilde{\varepsilon}$, where the scalar parameter θ determines the magnitude of this impact. This means that a shock to the disturbance of region/observation i will have a direct impact arising from $\tilde{\varepsilon}$ and indirect impact only on neighboring observations $\theta W \tilde{\varepsilon}$. Feedback effects are ruled out since the matrix W contains zeros on the diagonal. Diffusion to higher-order neighbors is also ruled out by the lack of terms involving higher powers of the matrix W .

77.2.2 Spatial Lag of X Models

Two other spatial regression models that LeSage and Pace (2009) label a spatial lag of X (SLX) and spatial Durbin error model (SDEM) models are shown in Eqs. (77.5) and (77.6):

$$y = \alpha I_n + X\beta + WX\theta + \varepsilon \quad (77.5)$$

$$y = \alpha I_n + X\beta + WX\theta + u, \quad u = (I_n - \rho W)^{-1}\varepsilon \quad (77.6)$$

These models allows for *local spatial spillovers* which can be directly calculated using the coefficients θ . This can be seen by considering the matrix expression for the partial derivative of y with respect to changes in the r th explanatory variable shown in Eq. (77.7):

$$\partial y / \partial x^{r'} = (I_n \beta_r + W \theta_r) \quad (77.7)$$

An implication of this model is that changes in the characteristics of the r th variable for a single observation i can potentially impact all observations in the vector y . Since we could consider changes in each observation $i = 1, \dots, n$, we have an $n \times n$ matrix of responses in y . This observation was made by Kim et al. (2003) as well as Kelejian et al. (2006). LeSage and Pace (2009) made the point that an $n \times n$ matrix of partial derivative responses poses a real problem for reporting estimates of the marginal impacts from these models. Consider that an application involving the 3,109 counties in the lower 48 US states would produce a $3,109 \times 3,109$ matrix of responses for each of the r explanatory variables (although many of these responses would equal 0 if W is sparse). A further issue is that we would like to have measures of dispersion for our marginal effects estimates that would allow us to draw inferences regarding the statistical significance of these.

A solution to these two issues was proposed by LeSage and Pace (2009), who suggested using the average of the main diagonal elements of the $n \times n$ matrix ($I_n \beta_r + W \theta_r$) in Eq. (77.7) to produce a scalar summary measure of the direct effects. In addition, they proposed using an average of the (cumulated) off-diagonal elements as a scalar summary of the cumulative indirect effects. LeSage and Pace (2009) also provide an approach to calculating measures of dispersion for these scalar summary estimates, specifically standard deviations that can be used to construct t -statistics. This allows for standard regression-based interpretation of the marginal effects estimates from spatial regression models.

For the case of the SLX and SDEM models, these calculations simplify considerably. Since the main diagonal of the matrix W contains zeros, and the rows of the matrix W sum to one, this leads to the simple conclusion that the coefficient β_r reflects direct effects while θ_r captures spatial spillovers. The t -statistics reported by standard regression software algorithms should also provide a valid basis for

inference regarding the statistical significance of these effects estimates. The SDEM model allows for these same local spillovers with regard to the explanatory variables but also models global spatial diffusion of shocks that arise in the disturbance structure of the model, using the spatial autoregressive process $u = \rho Wu + \varepsilon$. To produce valid t -statistics, software algorithms for estimating spatial autoregressive error models would be required. Of course, one could also produce a model that relies on the spatial moving average process for the disturbances if there was interest in modeling local spatial diffusion of shocks that arise in the disturbances.

For the SDEM and SLX models, the coefficients in the vector θ represent local spillovers, since there is an impact only on immediately neighboring observations. We note that estimates from these two models should be similar, but in the face of spatial dependence in the disturbances, SDEM model estimates should be more efficient. The partial derivative expressions for both models are the same, but improved efficiency for the case of the SDEM model could impact inferences regarding significance of the direct and indirect effects estimates. Pace and LeSage (2008) provide a *Hausman test* that could be used to test for equality of the SLX and SDEM model coefficients. An absence of equality for these two sets of coefficients may provide evidence against the SDEM model in favor of a spatial lag variant such as the SAR or SDM models (see LeSage and Pace (2009)).

In these cases, interpretation of the coefficient estimates for β as a slope indicating how changes in X produce changes in y (on average over the sample) are valid, as in linear regression. Interpretation of the coefficient θ as reflecting how changes in the average neighboring characteristics impact y is also valid. Returning to our example where y reflects property values and X characteristics, the coefficient θ measures how changes in neighboring properties' characteristics impact the value of a typical property (on average over the sample). It is important to note that in this case there are no feedback effects like those described for the case of the SAR and SDM models. It is also true that the impacts arising from changes in properties' characteristics are restricted to fall only on neighboring properties, which is consistent with our definition of local spillovers.

77.2.3 Spatial Lag of y Models

The SAR and SDM models are shown in Eqs. (77.8) and (77.9), where y is an $n \times 1$ vector of outcomes and X is an $n \times k$ matrix of explanatory variables (excluding the constant term) with associated $k \times 1$ parameter vector β . The $n \times k$ matrix WX has been labeled a *spatial lag* of the explanatory variables and represents a linear combination of characteristics from neighboring regions/entities, with associated parameters γ . Similarly, the $n \times 1$ vector Wy is a spatial lag of the dependent variable, reflecting a linear combination of neighboring region values for the dependent variable. The intercept coefficient is α and I_n is an $n \times 1$ vector of ones. It is typically assumed that ε is normally distributed and obeys the Gauss-Markov assumptions:

$$y = \alpha \iota_n + \rho W y + X\beta + \varepsilon \quad (77.8)$$

$$y = \alpha \iota_n + \rho W y + X\beta + W X \gamma + \varepsilon \quad (77.9)$$

An examination of the data generating process for these models shown in Eqs. (77.10) and (77.11) makes it clear that they reflect a nonlinear relationship between y and the right-hand side terms ι_n, X and ε :

$$y = (I_n - \rho W)^{-1} (\alpha \iota_n + X\beta + \varepsilon) \quad (77.10)$$

$$y = (I_n - \rho W)^{-1} (\alpha \iota_n + X\beta + W X \gamma + \varepsilon) \quad (77.11)$$

As already noted, the inverse $(I_n - \rho W)^{-1}$ can be expressed as an infinite sequence: $I_n + \rho W + \rho^2 W^2 + \rho^3 W^3 + \dots$, and the matrix product WX reflects a linear combination of the explanatory variables from neighboring regions. The matrix product W^2X creates a linear combination involving neighbors to the neighboring regions, or what are sometimes called second-order neighbors. As noted, diagonal elements of W^2 are not zero, since regions will by definition be neighbors to their neighbors. This means that feedback effects are present in models involving higher-order neighbors. As a concrete example, consider a homeowner A improving her property (the matrix X might contain property characteristics and the vector y property values), which has a beneficial spillover effect on the value of a neighboring property B (as well as on property A of course). However, in the SAR and SDM models, a change in the value of neighboring property B will feedback on the value of property A since the vector Wy included in the model would contain the increased property value of the neighbor. A similar statement regarding nonzero diagonal elements applies to higher-order matrix products such as W^3X which reflects a linear combination of neighbors to the neighbors, to the neighbors, and so on for higher powers. This is consistent with our definition of global spillovers. From the model statements in Eqs. (77.8) and (77.9), we can see the source of feedback works through impacts on the price of neighboring properties.

One implication of the nonlinear relationship in the SAR and SDM models between y and X is that the coefficients α, β , and γ cannot be interpreted as if they reflect linear regression slope estimates. This type of situation arises in a number of other nonlinear regression models such as probit and Tobit. The econometrics literature interprets coefficients from these models using marginal effects that reflect partial derivatives indicating how changes in each explanatory variable impact (or effect) the expected y outcomes. It should be clear that we need to take a similar approach for our nonlinear relationships between y and X in the SAR and SDM models.

That approach is based on the $n \times n$ matrices of partial derivatives for these models, shown in Eq. (77.12) for the SAR and Eq. (77.13) for the SDM:

$$\partial y / \partial x^{r'} = (I_n - \rho W)^{-1} I_n \beta_r \quad (77.12)$$

$$\partial y / \partial x^r = (I_n - \rho W)^{-1} (I_n \beta_r + W \gamma) \quad (77.13)$$

As in the case of the SLX and SDEM models, there is a need to produce scalar summary measures of the $n \times n$ matrices of partial derivatives. We can do this following the suggestion of LeSage and Pace (2009), using the mean of the main diagonal elements of the $n \times n$ matrices in Eqs. (77.12) and (77.13) to produce a scalar summary of the *direct effects*. These show how changes in the r th explanatory variable for the i th region impact the i th region's dependent variable, for $i = 1, \dots, n$. Using the mean of these n , different values produces a scalar summary that can be interpreted as representing how a change in the r th explanatory variable in the typical or representative region impacts outcomes y for the typical region.

For the case of linear regression relationships where $\rho = 0$ so observations are assumed to be independent, the partial derivatives $\partial y_i / \partial x_i^r = \beta_r, i = 1, \dots, n$. This means there is no need to produce a scalar summary estimate based on the mean of these n different partial derivatives in linear regression models.

Indirect effects representing the impact on the j th region outcomes y_j from a change in the r th explanatory variable from the i th region are captured by the off-diagonal elements of the $n \times n$ matrices in Eqs. (77.12) and (77.13). Specifically, the elements in the i th row show $\partial y_i / \partial x_j^r, j \neq i, j = 1, \dots, n$, reflecting how changes in each of the $j \neq i, j = 1, \dots, n$ or other regions' r th explanatory variable impact outcomes in the i th region. Since these are partial derivatives, the off-diagonal elements in the i th column represent $\partial y_j / \partial x_i^r, j \neq i, j = 1, \dots, n$ or how changes in the r th variable in region i impact outcomes across all regions $j \neq i$. LeSage and Pace (2009) suggest using the mean of the sum of off-diagonal elements from each row to produce a scalar summary measure of *cumulative indirect effects* or spatial spillovers. They note that the numerical magnitude of these equals the mean of the off-diagonal elements from each column, which would produce the same scalar summary measure of cumulative indirect effects.

In the case of linear regression relationships where $\rho = 0$ and observations are independent, the partial derivatives $\partial y_i / \partial x_i^r = 0$, so there are no spatial spillovers or indirect effects.

Elhorst (2010) notes that for the SAR model all coefficients $\beta^r, r = 1, \dots, k$ are multiplied by the same matrix $(I_n - \rho W)^{-1}$ (see Eq. (77.12)), so this model implies that the ratio between the indirect and direct effects is the same for every explanatory variable. The magnitude determining this ratio will depend only on the spatial dependence parameter ρ and the spatial weight matrix W . Another implication of the relationship between direct and indirect effects for the SAR model is that the sign of spillovers (indirect effects) and direct effects must be the same for the r th variable and will be determined by the sign of the coefficient estimate β_r . Of course, the signs can vary across the different explanatory variables in the model.

The relationship between direct and indirect effects in the SDM (and SLX and SDEM) models is not subject to these constraints. For these models, we can have positive (or negative) direct effects associated with negative (or positive)

indirect effects for the r th variable, so that spillover impacts might work in the opposite direction of direct impacts arising from changes in each explanatory variable.

To illustrate these issues, consider a cross-sectional model of state-level cigarette sales (an $n \times 1$ vector y) as a function of two explanatory variables, state-level cigarette taxes and income and a constant (an $n \times 3$ matrix X). In the SDM, SLX, or SDEM models, an increase in state i 's tax on cigarettes may generate a direct effect that decreases sales of cigarettes in state i (the dependent variable y_i) (on average across all states) but increases cigarette sales in neighboring states (y_j , a positive cumulative indirect effect averaged over all states), due to the presence of cross-border shopping for cigarettes in lower-tax neighboring states. This would not be possible in the SAR model, where the direct and indirect effects must have the same sign. For the SAR model, it would also be the case that the relative size of the direct and indirect effects for a change in tax rates (one explanatory variable in the model) must be the same as the relative size of the direct and indirect effects arising from a change in state income (another explanatory variable in the model). Practitioners should be aware of these facets of the SAR, SDM, SLX, and SDEM models to make an appropriate choice of model specification for the particular situation being modeled.

77.2.4 Measures of Dispersion for the Effects Estimates

In addition to calculating point estimates for the direct, indirect, and total effects associated with changing explanatory variables in the various types of spatial regression models, we require measures of dispersion for inference. Given an estimate of the variance (standard deviation) for the scalar summary point estimates, we can test hypotheses regarding the significance of the various types of effects for each of the explanatory variables used in the model.

In the case of SEM, SLX, and SDEM models, traditional standard deviations and asymptotic t -statistics based on maximum likelihood (or Bayesian) estimation would provide a valid basis for inference regarding the effects estimates. Of course for the SEM, there are only direct effects, since this model assumes indirect (spillover) effects are zero. For the SLX model, use of the t -statistic for the spatially lagged explanatory variables based on standard least-squares estimates would provide a valid basis for inference regarding significance of the direct and indirect or spillover effects for each variable. Coefficients associated with the explanatory variables X represent direct effects while those associated with WX are the indirect effects. If one were interested in significance of the total effect, this would require calculating a t -statistic for a distribution reflecting the sum of the coefficients on both X and WX , and this is typically not a part of standard regression software.

In the case of the SDEM model, asymptotic t -statistics from maximum likelihood, Bayesian, or instrumental variables estimation on the variables X and WX would provide a valid basis for inference regarding direct and indirect effects respectively. As noted for the case of the SLX model, measures of dispersion for

the total effects would require consideration of dispersion for a distribution based on the sum of the direct plus indirect effects.

Inference for the SAR and SDM models regarding the direct, indirect, and total effects is more involved. The point estimates for β do not directly measure the partial derivative effects, so reported t -statistics cannot be used as a basis for inference.

For maximum likelihood SAR and SDM estimates, measures of dispersion can be constructed by simulating values for the parameters from the estimated variance-covariance matrix. (See Pace ► [Chap. 78, “Maximum Likelihood Estimation”](#) for a discussion of the form taken by the variance-covariance matrix.) These simulated values (say 1,000 values) for the model parameters ρ, β can be used in Eqs. (77.12) or (77.13) for the SAR or SDM models to produce 1,000 values for the scalar summary effects estimates. Taking the median of these simulated summary measures would provide a point estimate for the summary measures. We take the median in this situation since the summary measures are not symmetrically distributed. For example, the total effect for the SAR model equals $\tilde{\beta}_r(1 - \tilde{\rho})^{-1}$. Even if $\tilde{\beta}$ and $\tilde{\rho}$ follow normal distributions, the total impact statistic will not follow a normal distribution and the median can provide a better measure of the central tendency in this situation. Similarly, one can use other measures of dispersion such as the scaled median absolute deviation (1.48 MAD) to provide a measure of dispersion of the median. Use of the median and the scaled MAD allows inference on the direct, indirect, and total impacts. LeSage and Pace (2009) provide a computationally efficient approach to processing the draws for the parameters to produce empirical distributions for the direct, indirect, and total effects estimates. This computationally efficient approach has been implemented in the Spatial Econometrics Toolbox for MATLAB (LeSage 1999), the R-language routines (Bivand and Albrecht 2000), and the Stata package for spatial regression (Drukker et al. 2001).

For SAR and SDM models estimated using Bayesian Markov chain Monte Carlo as described in ► [Chap. 79, “Bayesian MCMC Estimation,”](#) the draws used to produce point estimates can be used in place of parameter draws made based on the maximum likelihood estimate of the variance-covariance matrix. These draws can be used in conjunction with the computationally efficient formulas from LeSage and Pace (2009).

SAR and SDM models estimated using instrumental variable methods can produce draws using an asymptotic approximation to the variance-covariance matrix. (See Prucha and Jennish ► [Chap. 80, “Instrumental Variables/Method of Moments Estimation”](#) for details concerning this type of estimation and the variance-covariance matrix.) These draws can be used in conjunction with the computationally efficient formulas from LeSage and Pace (2009).

77.2.5 Partitioning Global Effects Estimates Over Space

For effects estimates involving local spillovers, these fall only on immediately neighboring regions and do not generate feedback effects. In contrast, we have used

global spillovers to reflect situations where there are feedback effects and spillovers fall on neighbors to neighbors, neighbors to neighbors to neighbors, and so on for higher-order neighbors.

For cases involving global spillovers, we might be interested in the pattern of decay of influence in the direct and indirect effects as we consider impacts on first-, second-, third-, and higher-order neighbors.

If we consider the matrix inverse in Eqs. (77.12) and (77.13), $(I_n - \rho W)^{-1}$ can be expressed as an infinite series: $I_n + \rho W + \rho^2 W^2 + \rho^3 W^3 + \dots$; it should be clear that we can partition the effects estimates by order of the matrix W . For example, expressions for the impacts on second-order neighbors (those involving W^2) based on the $n \times n$ matrix of partial derivatives are shown in Eqs. (77.14) and (77.15) for the SAR and SDM models:

$$\frac{\partial y}{\partial x^{r'}} = \rho^2 W^2 \beta_r \quad (77.14)$$

$$\frac{\partial y}{\partial x^{r'}} = \rho^2 W^2 (I_n \beta_r + W \gamma) \quad (77.15)$$

Scalar summaries can be constructed for these two $n \times n$ matrices using diagonal and off-diagonal elements as described in Sect. 77.2.3. Indirect effects (scalar summaries) constructed from the off-diagonal elements of these matrices would show the (average) spillover impacts falling only on second-order neighbors. Direct effects (scalar summaries) based on the main diagonal elements of these matrices would be feedback effects on the own region arising from second-order neighbors.

77.3 Applications of Spatial Regression Models

Applications of the various spatial regression models that have appeared in the literature illustrating correct interpretation are presented in this section.

77.3.1 Spatial Error Models

Bell and Bockstael (2000) use a spatial error model in an application involving land parcels. The hedonic model uses a sample of 1,000 residential sales transactions involving the assessed value of improved land parcels on which homes were built in the previous year. The goal is to examine which property and environmental/locational characteristics influence the new home sales prices. Explanatory variables include things like the (log) of assessed value of improvement in the land parcel, size of the parcel, travel distance to the two nearest major metropolitan areas, planned infrastructure characteristics such as sewers, public and private open space surrounding the homes, and measures of density and land use types for surrounding areas.

Estimates for the various parcel characteristics from the SEM can be interpreted as the (average) direct impact on the selling price arising from changes in the

various characteristics, just as in ordinary least squares. This simplicity arises from the implicit assumption of no spillovers, only contagion arising from shocks to the disturbances.

Bell and Bockstael (2000) rely on three different (row-normalized) contiguity-based weight matrices that assign values of 0 or 1 to neighboring observations that are within 200, 400, and 600 m distances of each observation. Estimation results from models using these three matrices are compared to a fourth (row-normalized) matrix where inverse distance-based weights were assigned to each neighboring observation within 600 m. They rely on a spatial error model (SEM) shown in Eq. (77.16) and compare estimates from least squares, maximum likelihood, and generalized moments, with the latter two sets of estimates constructed using the four different types of spatial weight matrices:

$$y = X\beta + u \quad (77.16)$$

$$u = \rho Wu + \varepsilon \quad (77.17)$$

An implicit assumption of the SEM model is that there are no omitted variables that are correlated with included explanatory variables. The presence of omitted variables leads to an SDM specification (LeSage and Pace 2009, pp. 27–28). A simple test for the appropriateness of the SEM would be to compare least-squares (OLS) estimates for the coefficients β to those from the SEM since we know these should theoretically be the same. This should be true irrespective of the spatial weight matrix used, since changes in the spatial weight specification could lead to changes in measures of dispersion (e.g., t -statistics), but not significant differences in the coefficients β . Any (significant) differences between estimates from these two models should reside in the measures of dispersion which would have an impact on the t -statistics, but not the point estimates for β .

Pace and LeSage (2008) use this idea to develop a formal Hausman specification test for significant differences between OLS and SEM estimates for β . Intuitively, significant differences in OLS and SEM estimates for β point to model misspecification that should lead us to reject the SEM model as an appropriate choice.

Examining estimates from Table 2 in Bell and Bockstael (2000), it is clear that five of the ten coefficients β from OLS estimation versus maximum likelihood estimation of the SEM model differ by more than 1.67 standard deviations. Table 77.1 presents their OLS and maximum likelihood SEM estimates constructed using an inverse distance weight matrix based on a 600 m cutoff (from Table 2 in their paper), along with standard errors and a t -test for significant differences between these.

There is one coefficient where the SEM estimate is 2.8 standard deviations away from the OLS, two cases where the two sets of estimates are 1.99 standard deviations apart, and two more that are different using the 90 % level of significance. Of the ten coefficients, five are likely to be significantly different, suggesting the SEM model represents a misspecification. This type of empirical comparison of OLS and SEM estimates should be a standard part of empirical studies using the SEM specification.

Table 77.1 Bell and Bockstaal (2000) OLS and maximum likelihood SEM estimates

	OLS $\hat{\beta}_o (\hat{\sigma}_{\beta_o})$	ML $\hat{\beta}_{ml} (\hat{\sigma}_{\beta_{ml}})$	<i>t</i> -statistic (<i>t</i> -probability) $H_o : \hat{\beta}_o = \hat{\beta}_{ml}$
Intercept	4.7332 (0.2047)	5.1725 (0.2204)	1.9932 (0.0465)
LIV	0.6926 (0.0124)	0.6537 (0.0135)	2.8815 (0.0040)
LLT	0.0079 (0.0052)	0.0002 (0.0052)	1.4808 (0.1390)
LDC	-0.1494 (0.0195)	-0.1774 (0.0245)	1.1429 (0.2534)
LBA	-0.0453 (0.0114)	-0.0169 (0.0156)	1.8205 (0.0690)
POPN	-0.0493 (0.0408)	-0.0149 (0.0414)	0.8309 (0.4062)
PNAT	0.0799 (0.0177)	0.0586 (0.0212)	1.0047 (0.3153)
PDEV	0.0677 (0.0180)	0.0253 (0.0253)	1.6759 (0.0941)
PLOW	-0.0166 (0.0194)	-0.0374 (0.0224)	0.9286 (0.3533)
PSEW	-0.1187 (0.0173)	-0.0828 (0.0180)	1.9944 (0.0464)

77.3.2 SLX and SDEM Models

An example that uses the SDEM model is LeSage and Ha (2012), who study the impact of migration on county-level social capital. Social capital is sometimes thought to be embodied in the *structure and relations between people*, which would suggest that social capital is *place based*, since the structure and relations between people exist at some particular location in space. Other definitions that emphasize the place-based nature of social capital rely on the concept of *associational density* which measures the number of civic and social organizations (per capita), again at some particular location in space. Others have emphasized a view of social capital that focuses on the structure of relations between people, using trust as a measure of the strength between individuals, with trust purported to promote positive economic performance of regional economies. When people move, they may take their trusting attitudes with them, so the case for social capital as entirely place based may not be the entire story.

Their SDEM model takes the form in Eq. (77.18), where W_n and W_f represent migration-weighted spatial weight matrices. The matrix W_n identifies neighboring counties within 40 miles and assigns relative weights to these based on in-migration magnitudes. The matrix W_f identifies neighboring counties more than 40 miles away from each county that provide in-migration to each county i in the sample and weighs these according to in-migration magnitudes. The matrix V used to model dependence in the model disturbances was a spatial contiguity weight matrix, with equal weights assigned to all contiguous counties:

$$\begin{aligned} y &= X\beta + W_n X\theta + W_f X\gamma + u \\ u &= \rho Vu + \varepsilon \end{aligned} \tag{77.18}$$

Pace and Zhu (2012) point out that a desirable aspect of the model in Eq. (77.18) is that dependence in the disturbances is modeled separately from spillovers, which is not the case for the SAR and SDM models. For the SAR model, the dependence

structure for the disturbances is restricted to be the same as that for the *mean model*, which can be seen from $y = (I_n - \rho W)^{-1}X\beta + (I_n - \rho W)^{-1}\varepsilon$. This implies that the expectation for y is a function of ρ and W , $E(y) = (I_n - \rho W)^{-1}X\beta$, and the disturbance covariance, $\Omega = \sigma^2[(I_n - \rho W)^{-1}(I_n - \rho W')^{-1}]$, takes the same functional form. An implication of this is that misspecification in either the disturbances or mean model will contaminate the other part of the model.

The SDEM model in Eq. (77.18) allows *separation* of the (local) spillover impacts on county-level social capital levels arising from changes in population characteristics of nearby counties (providing in-migrants to each county in the sample) versus (local) spillover impacts that arise from changes in population characteristics of far away (outside the region) counties (providing in-migrants to each county) and global contagion impacts from shocks to the errors. An important point is that local spillovers need not be defined as those involving only nearby counties; we can consider spillovers in far away counties as well with this type of model. The intuition behind local spillovers is that the spatial extent of these falls only on first-order neighbors, and a model of local spillovers exhibits no feedback effects.

In terms of the application to migration impacts on social capital levels, the model allows us to consider how a change in educational attainment levels of population in counties within the region impact social capital and how similar changes in educational attainment for population in counties outside the region (providing in-migrants to each county) impact levels of social capital. As is conventional in regression models, the coefficient estimates θ and γ provide answers to this type of question by averaging over all observations/counties in the sample, so we interpret these to represent impacts on the typical county.

This type of model allows us to focus on whether there are important differences in the magnitude of impact associated with in-migration from within and outside the region. Are some changes in characteristics of in-migrants from nearby counties significant/insignificant while the same characteristics of in-migrants from outside the region are insignificant/significant?

A specification test to rule out the presence of omitted explanatory variables that are correlated with included variables should be employed here. Theoretically, we know that SLX and SDEM model estimates for the coefficients θ and γ should be the same, with differences residing in the t -statistics. If omitted variables that are correlated with the included explanatory variables exist, there would be a significant difference between SLX and SDEM estimates for β . LeSage and Ha (2012) provide a comparison of SLX and SDEM estimates in their Table 2, where no significant differences exist.

77.3.3 SAR and SDM Models

Kirby and LeSage (2009) use an SDM specification to consider changes in the (logged) number of workers in the US census tracts with commuting times exceeding 45 min one way, between 1990 and 2000. They motivate their investigation by

noting that the percentage of the US workers with these long commute times in 1990 was 12.5 %, compared to 15.4 % in 2000, an increase of more than 10 %. Spillover impacts from an increase in commuters traveling long distances to work would seem global in nature, since the congestion effects of more travelers on one segment of a metropolitan area roadway network impact travel times of other travelers on the entire network. Additionally, feedback effects seem likely since congestion arising from commuting decisions by workers in one tract will spillover to neighboring tracts, which in turn create congestion feedback to the own tract. Intuitively, congestion on roadways does not obey census tract boundaries, but spills over to the neighboring tracts. Any traffic backups on a road segment in neighboring tracts are likely to cross tract boundaries back into the own-tract roadway.

The SDM model in Eq. (77.19) is used to examine factors (X) that explain tract-level variation in the (logged) number of workers with long commute times (y), and the model includes these same characteristics of neighboring census tracts (WX):

$$y = \rho Wy + \alpha_{ln} + X\beta + WX\theta + \varepsilon \quad (77.19)$$

The motivation for these explanatory variables in the context of modeling long commuting times is that socioeconomic demographic characteristics of persons living in neighboring census tracts should represent important explanatory variables. For example, if there are a large number of retired persons living in neighboring tracts, this should result in less congestion during commute-to-work times.

They consider changes in both direct and indirect effects for models estimated using cross-sectional samples of all census tracts in the lower 48 states for the 1990 and 2000 periods. In the presence of significant indirect (spillover) effects, past studies that ignore these will produce biased and inconsistent estimates of how socioeconomic demographic characteristics impact commuting times for population living in the tracts. The analysis considers suites of variables that reflect *location decisions of households* (e.g., housing tenure, modes of transportation, residence versus work locations in cities versus suburbs), *age and gender and income distribution of resident population* (e.g., the number of persons in various age and gender categories, household income, and educational attainment), and *geographical characteristics of the tracts* including such things as public transport use, land and water area, and highway lane miles reflecting supply side considerations.

Based on a comparison of direct, indirect, and total effects estimates from the 1990 and 2000 models, they conclude that the suite of variables reflecting the age and gender distribution of population in the tracts represents the primary explanation for changes in the number of workers with long commute times between 1990 and 2000. This is in contrast to other studies that emphasize a rapid rise in household income leading to the desire for larger homes located farther from central business districts (Gordon et al. 2009). The spillover impacts of the number of employed females in the 1990 model was positive, suggesting that more employed females in a tract produced an increase in long commute times for

neighboring tract commuters. In contrast, for the 2000 relationships, spillovers associated with employed females were negative, so that more employed females in a tract reduced long commute times for workers located in neighboring tracts. Opposite signs were found for changes in spillovers associated with employed males, providing results that are consistent with observations made by others regarding changes in age and gender impacts on commuting behavior (Crane and Chatman 2004).

77.4 Conclusion

An important motivation for use of spatial regression models is that they allow regional scientists to quantify spatial spillovers, which represent a major theme in regional science. Past studies using spatial regression models frequently interpreted the model estimates incorrectly, which tended to obscure this valuable aspect of spatial regression models.

We draw on past work by LeSage and Pace (2009) that formally defines spatial spillovers in the context of spatial regression models. An important distinction in applied modeling can be drawn between local and global spillovers. Global spillovers arise when changes in a characteristic of one region impact outcomes in more than just immediately neighboring regions. Global impacts fall on neighbors to the neighbors, neighbors to the neighbors to the neighbors, and so on. In contrast, local spillovers represent a situation where the impacts fall only on nearby or immediate neighbors, dying out before they impact regions that are neighbors to the neighbors. Another distinction between local and global spillovers is that feedback effects arise in the case of global spillovers, but not in the case of local spillovers.

Different types of spatial regression models should be used to model local versus global spillovers. Reasoning about the local versus global nature of spillovers in particular applications seems a useful approach to selecting a model specification from the family of spatial regression models. Past work by practitioners may have placed too much emphasis on statistical tests based on model fit to distinguish between alternative specifications from the family of spatial regression models. It might be worthwhile to devote more effort to reasoning about the likely nature of spatial spillovers in particular applied situations.

References

- Allers M, Elhorst JP (2005) Tax mimicking and yardstick competition among local governments in the Netherlands. *Int Tax Public Financ* 12(4):493–513
- Anselin L (1988) Spatial econometrics: methods and models. Kluwer, Dordrecht
- Anselin L (2003) Spatial externalities, spatial multipliers and spatial econometrics. *Int Reg Sci Rev* 26(2):153–166
- Bell KP, Bockstaal NE (2000) Applying the generalized-moments estimation approach to spatial problems involving microlevel data. *Rev Econ Stat* 87(1):72–82
- Bivand R, Albrecht G (2000) Implementing functions for spatial statistical analysis using the R language. *J Geogr Syst* 2(3):307–317

- Crane R, Chatman D (2004) Traffic and sprawl: evidence from U.S. commuting, 1985–1997. *Plan Mark* 6(3):14–22
- Deskins J, Hill B (2010) Have state tax interdependencies changed over time? *Public Financ Rev* 38(2):244–270
- Drukner DM, Prucha I, Raciborski R (2001) A command for estimating spatial-autoregressive models with spatial-autoregressive disturbances and additional endogenous variables. *Stata J* 1(3):1–13
- Elhorst JP (2010) Applied spatial econometrics: raising the bar'. *Spat Econ Anal* 5(1):9–28
- Fingleton B (2001) Theoretical economic geography and spatial econometrics: dynamic perspectives. *J Econ Geogr* 1(2):201–225
- Gordon P, Lee B, Richardson HW (2009) Commuting trends in U.S. cities in the 1990s. *J Plan Educ Res* 29(1):78–89
- Haining R (1990) Spatial data analysis in the social and environmental sciences. Cambridge University Press, Cambridge
- Kelejian HH, Tavlas GS, Hondroyiannis G (2006) A spatial modeling approach to contagion among emerging economies. *Open Econ Rev* 17(4/5):423–442
- Kim CW, Phipps TT, Anselin L (2003) Measuring the benefits of air quality improvement: a spatial hedonic approach. *J Environ Econ Manag* 45(1):24–39
- Kirby DK, LeSage JP (2009) Changes in commuting to work times over the 1990 to 2000 period. *Reg Sci Urban Econ* 39(4):460–471
- LeSage JP (1999) The theory and practice of spatial econometrics, a manual to accompany the spatial econometrics toolbox. Freely available at: www.spatialeconometrics.com
- LeSage JP, Ha C (2012) The impact of migration on social capital – do migrants take their bowling balls with them? *Growth Change* 43(1):1–26
- LeSage JP, Pace RK (2009) Introduction to spatial econometrics. Taylor Francis/CRC Press, Boca Raton
- Ord JK (1975) Estimation methods for models of spatial interaction. *J Am Stat Assoc* 70(3):120–126
- Pace RK, LeSage JP (2008) A spatial Hausman test. *Econ Lett* 101(3):282–284
- Pace RK, Zhu S (2012) Separable spatial modeling of spillovers and disturbances. *J Geogr Syst* 14(1):75–90

R. Kelley Pace

Contents

78.1	Introduction	1553
78.2	Likelihoods	1554
78.2.1	Likelihood with Continuous y and Normal Disturbances	1554
78.2.2	Likelihood with Binomial y and Normal Disturbances	1555
78.3	Inference and Estimation	1556
78.3.1	Simplest Approach to Likelihood Estimation and Inference	1556
78.3.2	Variance-Covariance Matrix Approach to Likelihood Inference	1557
78.4	Spatial Error Model Example	1559
78.5	Computational Details	1562
78.5.1	Concentrated Log-Likelihood	1562
78.5.2	Sparsity	1563
78.5.3	Log-Determinant Calculations	1564
78.6	Conclusions	1568
	References	1568

Abstract

Maximum likelihood estimation has been the standard method employed for estimating spatial econometric models. This chapter introduces these methods, examines the specific case of a spatial error model, and provides an example based on a large data set. In addition, the chapter sets forth various solutions to the computational difficulties that arise for large data sets.

R.K. Pace

Department of Finance, E.J. Ourso College of Business Administration, Louisiana State University, Baton Rouge, LA, USA
e-mail: kelley@pace.am

78.1 Introduction

At least since Ord (1975), maximum likelihood estimation has been the standard method employed for estimating spatial econometric models. Maximum likelihood methods work under the simple, but powerful, idea that if the observed data come from some *data-generating process* or DGP based on a distribution and constant parameters, then inverting the process enables estimation of the parameters conditional on the observed data. In other words, for an assumed DGP, which values of the parameters would be in accord with having observed these data? The likelihood function has the same mathematical form as a density or distribution function but a different interpretation and properties.

Maximum likelihood methods provide a coherent approach to estimation and inference with attractive statistical properties. In the absence of misspecification, in most settings, maximum likelihood estimates are efficient in large samples.

This chapter briefly presents some spatial likelihoods in Sect. 78.2, sets forth the maximum likelihood approach in Sect. 78.3, works through an example based on the spatial error model in Sect. 78.4, goes into many of the computational aspects of maximum likelihood in Sect. 78.5, and summarizes the key ideas in Sect. 78.6.

78.2 Likelihoods

Since every different distribution has its own likelihood function, there are an enormous number of likelihood functions that have been introduced in the literature. Since dependent variables could be continuous, discrete, or a combination of both (as in the case of Tobit), this implies a variety of likelihood functions. In addition, the way that dependence between observations is specified affects the form of the likelihood function.

We present two distinct types of likelihood, one for the case involving normally distributed dependent data in 2.1 and one for the case of binary dependent variable observations in 2.2 where implied (latent) disturbances follow a multivariate normal distribution.

78.2.1 Likelihood with Continuous y and Normal Disturbances

Regression, whether linear or nonlinear, is the most commonly performed statistical procedure in regional science. Specifically, the DGP in Eq. (78.1) relates a given n by k full-rank matrix X that contains n observations on k explanatory variables and an n by 1 vector of observations on the dependent variable y as a function of a k by 1 parameter vector β and an n by 1 vector of disturbances ε that follows a multivariate normal distribution. The disturbance distribution is shown

in Eq. (78.2), where $\Omega(\lambda)$ is the variance-covariance matrix that depends on a scalar parameter λ and a scalar variance parameter σ^2 :

$$y = f(\beta|X) + \varepsilon \quad (78.1)$$

$$\varepsilon \sim N(0, \sigma^2 \Omega(\lambda)) \quad (78.2)$$

For this DGP, the relevant log of the likelihood $L(\omega)$ appears in Eq. (78.3), where ω is a $k + 2$ by 1 parameter vector shown in Eq. (78.4) and r corresponds to the unadjusted errors in Eq. (78.5):

$$L(\omega) = -\frac{n}{2} \ln(2\pi\sigma^2) - \frac{1}{2} \ln |\Omega(\lambda)| - \frac{r' \Omega^{-1} r}{2\sigma^2} \quad (78.3)$$

$$\omega = [\beta' \quad \sigma^2 \quad \lambda]' \quad (78.4)$$

$$r = y - X\beta \quad (78.5)$$

If a ω^* maximizes $L(\omega)$, this vector becomes the maximum likelihood estimate, which we label $\tilde{\omega}$.

In practice, the main computational constraint is the computation of the log-determinant of the n -dimensional variance-covariance matrix. This issue will be discussed below in Sect. 78.5.

78.2.2 Likelihood with Binomial y and Normal Disturbances

For the case of spatial dependence in the disturbance structure, the model in Eq. (78.6) represents a general data-generating process (DGP) with multivariate normal disturbances given in Eq. (78.7). The n by 1 vector \underline{y} represents unobserved utility/profits, and y is an associated 0,1 vector reflecting the observed decisions as in Eq. (78.8). Like the typical binary probit model, the parameter σ^2 is set to 1 to achieve identification as shown in Eq. (78.9). The link between the latent \underline{y} and the observed y appears in Eq. (78.10): (Reference the ► Chap. 81, “Limited and Censored Dependent Variable Models” by Cara Wang on spatial probit models here.)

$$\underline{y} = X\beta + \varepsilon \quad (78.6)$$

$$\varepsilon \sim N(0, \sigma^2 \Omega(\lambda)) \quad (78.7)$$

$$y = 0, 1 \quad (78.8)$$

$$\sigma^2 = 1 \quad (78.9)$$

$$\underline{y}_i > 0 \rightarrow y_i = 1 \text{ otherwise } y_i = 0 \text{ for } i = 1 \dots n \quad (78.10)$$

The log-likelihood in Eq. (78.11) involves the multivariate joint normal cumulative density function given the pattern of observed binary outcomes y and observed explanatory variables X :

$$L(\omega) = \ln \Phi_n(\omega|y, X) \quad (78.11)$$

Although this is simple conceptually, in practice, this requires computing the integral of a truncated n -dimensional normal distribution. For large n , this is known to be a difficult computational problem (Phinikettos and Gandy 2011). However, it is computationally possible using the GHK (Geweke-Hajivassiliou-Keane) simulator which arose out of the work by Geweke (1991), Hajivassiliou and McFadden (1990), and Keane (1994). In a spatial context, Beron and Vijverberg (2004) were the first to estimate spatial models using this technique, and recently, Pace and LeSage (2011) and Zhu and Pace (forthcoming) have used the sparsity of Ω to greatly accelerate computing the likelihood.

78.3 Inference and Estimation

Given a likelihood, one can follow at least two main approaches in using it for estimation and inference: the *Bayesian* approach and the maximum likelihood approach. This section introduces a simple maximum likelihood approach in 3.1 and sets forth the conventional maximum likelihood approach to estimation and inference based on derivatives in 3.2.

78.3.1 Simplest Approach to Likelihood Estimation and Inference

Maximization could come about through derivatives, grid searches, or a number of other techniques that have been suggested in the literature. One can compare the maximum likelihood estimate $\tilde{\omega}$ to some restricted estimate ω_0 using the log of the *likelihood ratio* in Eq. (78.12). Under standard likelihood theory, the *deviance* given in Eq. (78.13) is distributed as χ^2 with degrees of freedom (df) equal to the number of restrictions as in Eq. (78.14):

$$\ln LR = L(\omega_0) - L(\tilde{\omega}) \quad (78.12)$$

$$\text{Deviance} = -2 \ln LR \quad (78.13)$$

$$\text{Deviance} \sim \chi^2(\text{df}) \quad (78.14)$$

As an example, consider ordinary least squares (OLS) where we rely on a single data sample. The (log) likelihood for this model is a restricted version of the general normal model in Eq. (78.2), with the restriction being that $\lambda = 0$. Testing for consistency of the sample with the OLS versus the general model could be viewed as a test of the restriction that $\lambda = 0$. If OLS estimation of β and σ^2 produced an estimated value of the log-likelihood function of -100.0 , while estimation of β , λ , and σ^2 using the general model yielded a value of -75.0 , the deviance would equal 50.0 . Under the null hypothesis that $\lambda = 0$, the deviance would have a χ^2 distribution with one degree of freedom. This has a critical value at the one percent level of 6.63 , making it very unlikely that the disturbances are independent.

A related statistic is the *signed root deviance* which equals the square root of the deviance times the sign of the corresponding parameter (Chen and Jennrich 1996). In the example above, if $\lambda > 0$, the signed root deviance would equal to $\sqrt{50}$ or 7.07 , and this quantity can be interpreted like a t -statistic.

78.3.2 Variance-Covariance Matrix Approach to Likelihood Inference

The most common approach to estimation uses optimization in conjunction with partial derivatives to derive first-order conditions and second derivatives to arrive at a variance-covariance matrix for the parameter estimates (Cramer 1986; Davidson and MacKinnon 2004).

Specifically, the partial derivatives of the log-likelihood with respect to the parameters are termed the *Fisher's score function* or *score* as in Eq. (78.15). In other words, these are the *gradients* of the log-likelihood:

$$g(\omega) = \frac{\partial L(\omega)}{\partial \omega} \quad (78.15)$$

The *Hessian* in Eq. (78.16) contains the second partial derivatives of the log-likelihood:

$$H(\omega) = \frac{\partial^2 L(\omega)}{\partial \omega \partial \omega'} \quad (78.16)$$

The negative of the Hessian evaluated at the maximum likelihood estimates shown in Eq. (78.17) has been labeled the *observed information matrix*, and the expected value of the negative of the Hessian shown in Eq. (78.18) is referred to as the *information matrix*:

$$\tilde{J}(\omega) = -H(\omega) \quad (78.17)$$

$$I(\omega) = -E(H(\omega)) \quad (78.18)$$

Assuming the correct model specification, large samples imply Eq. (78.19):

$$I(\omega) = \tilde{J}(\omega) \quad (78.19)$$

For both of these information matrices, the variance-covariance matrix applicable to the parameter estimates is simply the inverse of the respective information matrices as in Eqs. (78.20) and (78.21):

$$V(\omega, I(\omega)) = I(\omega)^{-1} \quad (78.20)$$

$$V(\omega, \tilde{J}(\omega)) = \tilde{J}(\omega)^{-1} \quad (78.21)$$

Given the true variance-covariance matrix $V(\omega)$ and true parameters ω , the implied distribution of the estimates is shown in Eq. (78.22):

$$\tilde{\omega} \sim N(\omega, V(\omega)) \quad (78.22)$$

Of course, we do not know the true parameters, so Eqs. (78.23) and (78.24) provide a feasible version of Eq. (78.22):

$$\tilde{\omega} \sim N(\tilde{\omega}, V(\tilde{\omega}, I(\tilde{\omega}))) \quad (78.23)$$

$$\tilde{\omega} \sim N(\tilde{\omega}, V(\tilde{\omega}, \tilde{J}(\tilde{\omega}))) \quad (78.24)$$

Intuitively, if the second derivatives are negative in sign and large in magnitude (especially for elements on the diagonal of the Hessian or expected Hessian), this means that the log-likelihood is decreasing quickly for points away from the maximum likelihood estimate. The inverse of the negative of these negative, large magnitude second derivatives would yield small, positive variances for the respective parameter estimates.

In terms of the derivative approach to estimation, the gradient and the Hessian enable use of *Newton–Raphson* optimization, with the typical iteration step used to move from an intermediate value at step i to a new intermediate value for step $i + 1$ shown in Eq. (78.25):

$$\omega^{(i+1)} = \omega^{(i)} - H(\omega^{(i)})^{-1} g(\omega^{(i)}) \quad (78.25)$$

As with all iterative procedures, updated values continue based on these steps until convergence, which is defined by a predefined criterion on how close the gradient $g(\omega^{(i)})$ or the adjustment $H(\omega^{(i)})^{-1} g(\omega^{(i)})$ should be to a vector of zeros.

Fisher scoring is an optimization technique particularly well suited to maximum likelihood. Its optimization step Eq. (78.26) is a variation on the Newton–Raphson step where $-H(\omega)$ is replaced by the information matrix $I(\omega)$.

The technique derives its name since the score vector $g(\omega)$ is pre-multiplied by the inverse of the information matrix:

$$\omega^{(i+1)} = \omega^{(i)} + I(\omega^{(i)})^{-1} g(\omega^{(i)}) \quad (78.26)$$

In summary, the derivative-based approach outlined in this section for optimizing the log-likelihood and producing estimates can be used for inference works in a wide variety of settings. The next section provides specifics for this approach in the case of a *spatial error model* (SEM) specification.

78.4 Spatial Error Model Example

In this section, we apply the approach outlined above to the spatial error model, whose DGP appears in Eqs. (78.27)–(78.29):

$$y = X\beta + \varepsilon \quad (78.27)$$

$$\varepsilon \sim N(0, \sigma^2 \Omega) \quad (78.28)$$

$$\Omega = (I_n - \lambda W)^{-2} \quad (78.29)$$

We further assume that the n by n nonnegative matrix W with zero main diagonal is symmetric and thus has all real eigenvalues and eigenvectors (see Ord (1975) for the information matrix when using nonsymmetric W). In addition, we assume that the maximum eigenvalue of W equals 1. This means that $\Omega(\lambda)$ is symmetric and positive definite when $\lambda \in (-v_{\min}^{-1}, 1)$, where v represents the n by 1 vector of eigenvalues from W . We also assume an exogenous n by k matrix of explanatory variable observations X . As before, β is a k by 1 parameter vector, and σ^2 is a positive scalar parameter.

The normal log-likelihood appears in Eq. (78.30) with previous definitions repeated in Eq. (78.31). The spatial dependence structure of the error model appears in Eq. (78.32), which leads to the determinant expressions in Eq. (78.33). These relations enter into the spatial error model log-likelihood shown in Eq. (78.34). The spatial error model leads to a sum-of-squared error term Q that is quadratic in the dependence parameter λ as shown in Eq. (78.35):

$$L(\omega) = -\frac{n}{2} \ln(2\pi\sigma^2) - \frac{1}{2} \ln |\Omega(\lambda)| - \frac{r' \Omega^{-1} r}{2\sigma^2} \quad (78.30)$$

$$\omega = [\beta' \quad \sigma^2 \quad \lambda]', \quad r = y - X\beta \quad (78.31)$$

$$\Omega^{-1} = A'A, \quad A = I_n - \lambda W, \quad Q = r' \Omega^{-1} r \quad (78.32)$$

$$\ln |\Omega(\omega)| = -\ln |\Omega(\omega)^{-1}| = -2 \ln |I_n - \lambda W| \quad (78.33)$$

$$L(\omega) = -\frac{n}{2} \ln(2\pi\sigma^2) + \ln |I_n - \lambda W| - \frac{Q}{2\sigma^2} \quad (78.34)$$

$$Q = r' A^2 r = r' r - 2\lambda r' W r + \lambda^2 r' W^2 r \quad (78.35)$$

The gradient or score vector $g(\omega)$ values for the spatial error model appear in Eq. (78.36):

$$g(\omega) = \begin{bmatrix} \frac{\partial L(\omega)}{\partial \beta} \\ \frac{\partial L(\omega)}{\partial \sigma^2} \\ \frac{\partial L(\omega)}{\partial \lambda} \end{bmatrix} = \begin{bmatrix} \sigma^{-2} X' A^2 r \\ -\frac{n}{2}\sigma^{-2} + \frac{1}{2} Q \sigma^{-4} \\ -\text{tr}(B) + \sigma^{-2} r' W A r \end{bmatrix} \quad (78.36)$$

where use is made of the definitions in Eqs. (78.37) and (78.38). Note, matrix functions involving the same square matrix such as WA^{-1} commute, and so, this also equals $A^{-1} W$. This property along with symmetric W helps simplify the expressions:

$$B = WA^{-1} = A^{-1}W \quad (78.37)$$

$$\alpha = -\text{tr}(B^2) \quad (78.38)$$

The expression for α in Eq. (78.38) follows Ord (1975), but there is an omitted minus sign in Ord (1975, p. 124) that was corrected here.

Taking partial derivatives of the gradient leads to the Hessian (H) in Eq. (78.39) and expected Hessian $E(H)$ in Eq. (78.40):

$$H = \sigma^{-2} \begin{bmatrix} -X' A^2 X & -\sigma^{-2} X' A^2 r & -2X' A W r \\ \frac{n}{2}\sigma^{-2} - Q\sigma^{-4} & -\sigma^{-2} r' W A r & \alpha\sigma^2 - r' W^2 r \end{bmatrix} \quad (78.39)$$

$$E(H) = \begin{bmatrix} -\sigma^{-2} X' A^2 X & 0_{kx1} & 0_{kx1} \\ 0_{1xk} & -\frac{n}{2} & -\sigma^{-2} \text{tr}(B) \\ 0_{1xk} & -\sigma^{-2} \text{tr}(B) & 2\alpha \end{bmatrix} \quad (78.40)$$

As before, this leads to the observed information matrix in Eq. (78.41) and information matrix in Eq. (78.42), with estimates and associated variance-covariance matrix given in Eqs. (78.43) and (78.44):

$$\tilde{J}(\omega) = -H(\omega) \quad (78.41)$$

$$I(\omega) = -E(H(\omega)) \quad (78.42)$$

$$\tilde{\omega} \sim N(\tilde{\omega}, V(\tilde{\omega}, I(\tilde{\omega}))) \quad (78.43)$$

$$\tilde{\omega} \sim N(\tilde{\omega}, V(\tilde{\omega}, \tilde{J}(\tilde{\omega}))) \quad (78.44)$$

As an illustrative example that compares these methods, we consider a model where the dependent variable was year 2000 logged median house prices in 62,226 US Census tracts and a set of explanatory variables taken from the 1990 Census. Specifically, the explanatory variables were median house age (Hage), employment (Employ), median years of education (Edu), median age of the population (Age), and number of households (HHs). All of these variables were logged. In addition, the model includes an intercept and spatially lagged versions of the explanatory variables. LeSage and Pace (2009) term this the *spatial Durbin error model* or SDEM. The coefficients on the explanatory variables themselves represent the *direct* effects, and the coefficients on the spatially lagged explanatory variables represent the *indirect* effects (LeSage and Pace 2009) (reference to the ▶ Chap. 77, “Interpreting Spatial Econometric Models” by LeSage and Pace on interpreting spatial regression models here).

The SDEM model offers several advantages. First, the estimates are easily interpreted. Second, in contrast to the usual lag of y models, the dependence parameter λ does not affect the conditional mean from the model. In the traditional lag of y models, an incorrect specification of the disturbances could affect the direct and indirect effects extracted from the model. Since the disturbance parameter in the SDEM is separated from the conditional mean, this is a case of a *separable* model as discussed by Pace and Zhu (2012). Third, the SDEM should arrive at the correct direct and indirect effects even with misspecification of the disturbances (with enough observations).

The estimates from the SDEM model are shown in the first columns of Table 78.1, with those for the SLX model in the last columns. All three methods of calculating t -statistics are reported for the SDEM model, with the column t_I representing information matrix results, t_H the Hessian, and t_{dev} the signed root deviances.

Estimates of the direct effects show the anticipated signs so that older housing reduces the expected future price, while employment, education, and age of the population (a proxy for wealth) increase the expected future price. In terms of the *local spillovers* or indirect effects, these all have the same signs as the direct effects, and so, the spillovers in this case all reinforce the direct effects. In other words, the total effects are all larger than the direct effects for this example.

From the standpoint of the three methods used for inference (likelihood ratio, observed information matrix, and information matrix), these all produce similar t -statistics on the explanatory variables. However, the t -statistics on $\tilde{\lambda}$ do vary. This arises partly because the distribution of $\tilde{\lambda}$ is not quadratic over all the domain of λ .

Table 78.1 *t*-Statistics across methods

Model	SDEM model				SLX model	
	$\tilde{\beta}$	t_I	t_H	t_{dev}	$\hat{\beta}$	t
Intercept	4.5451	35.435	35.1615	35.258	2.6050	36.081
HAge	-0.1023	-42.797	-42.6704	-42.470	-0.0961	-20.836
Employ	0.4427	67.488	66.9661	65.070	0.4621	38.572
Edu	0.8437	68.887	68.7295	67.090	0.8475	37.181
Age	0.5570	68.579	68.5792	67.195	0.5487	37.686
HHs	-0.4567	-67.391	-66.9431	-65.150	-0.4810	-39.662
W-HAge	-0.0098	-1.637	-1.6251	-1.637	0.1184	20.948
W-Employ	0.4387	26.032	25.5814	25.134	1.0133	63.578
W-Edu	0.7159	22.694	22.5603	22.335	1.3185	42.250
W-Age	0.3876	18.104	18.1041	18.060	0.3507	16.814
W-HHs	-0.4267	-24.228	-23.8796	-23.551	-0.9921	-58.515
$\tilde{\lambda}$	0.8710	402.791	471.1626	268.947	0.0000	0.000

(although it can be locally quadratic), and it may point to some misspecification. Much of this comes from fatter tails in the residuals than found in a normal distribution. In cases involving misspecification of the disturbances, maximum likelihood is consistent but not asymptotically efficient. Normal maximum likelihood applied to data with non-normal disturbances is termed *quasi-maximum likelihood*.

See Mardia and Marshall (1984), Anselin (1988), Griffith (1989), and Haining (1990) for more on the derivative approach to spatial model estimation and inference. See Burridge (2012) for the necessary derivatives for more general spatial models with normal disturbances.

78.5 Computational Details

Even though computational power has greatly increased since Ord's seminal article in 1975 on normal maximum likelihood for spatial models, a brute force approach to estimation for large n can still encounter difficulties. However, a number of techniques can greatly reduce the computational effort required in normal maximum likelihood. These include use of profile or concentrated log-likelihoods (5.1), sparse matrices (5.2), and various approaches to calculating the log-determinant term (5.3).

78.5.1 Concentrated Log-Likelihood

Beginning with the spatial error model log-likelihood in Eq. (78.45),

$$L(\omega) = -\frac{n}{2} \ln(2\pi\sigma^2) - \frac{1}{2} \ln |\Omega(\lambda)| - \frac{r' \Omega(\lambda)^{-1} r}{2\sigma^2} \quad (78.45)$$

substitution of the solutions to the first-order conditions ($g(\beta)$, $g(\sigma^2)$) given by Eqs. (78.46) and (78.47) into Eq. (78.45) yields $L_p(\lambda)$, a function of only the parameter λ :

$$\beta = (X' A^2 X)^{-1} X' A^2 y \quad (78.46)$$

$$\sigma^2 = \frac{1}{n} Q(\lambda) \quad (78.47)$$

$$L_p(\lambda) = \kappa + \ln |I_n - \lambda W| - \frac{n}{2} \ln(Q(\lambda)) \quad (78.48)$$

The univariate function $L_p(\lambda)$ in Eq. (78.48) is termed a *concentrated* or *profile* log-likelihood. The value of λ^* that maximizes $L_p(\lambda)$ also maximizes $L(\omega)$ given Eqs. (78.46) and (78.47). Since $L_p(\lambda)$ is univariate, it is easier to optimize. Note that the variance of λ^* implied solely by the concentrated log-likelihood second derivative with respect to λ does not match the variance of λ coming from the information matrix approach. However, given the variance of λ from the concentrated log-likelihood, a transformation can be made to obtain the exact value of the variance when following the information matrix approach (Davidson and MacKinnon 2004; LeSage and Pace 2009, p. 56–59).

78.5.2 Sparsity

In the SEM example, the variance-covariance matrix appears in Eq. (78.49), and the inverse of the variance-covariance matrix, known as the *precision matrix* which is labeled Ψ , appears in Eq. (78.50):

$$\Omega = (I_n - \lambda W)^{-2} = A^{-2} \quad (78.49)$$

$$\Psi = \Omega^{-1} = (I_n - \lambda W)^2 = A^2 \quad (78.50)$$

In time series analysis, an AR(1) process corresponds to a case where the equivalent of A is triangular with at most a single off-diagonal nonzero in each row. In that case of the n^2 elements making up the n by n matrix A , there are less than $2n$ nonzero elements, which can be expressed as a density of $2n/n^2 = 2/n$ nonzeros. For the MA(1) process, the same holds true for the equivalent of A^{-1} . In other words, for simple time series processes, either the variance-covariance matrix or the precision matrix contains many zero elements.

This is often true for spatial processes, since the most common spatial weight matrix W contains nonzero elements for regions/observations where borders of two regions touch. This leads to an average of approximately 6 nonzeros for each row of W or a density of $6/n$ elements. A nearest-neighbor-based W would

have a density of m/n , where m is the number of neighbors. A similar number exists for distance-based spatial weight matrices where a cutoff distance is used to assign zero values for regions beyond the cutoff. Matrices with a low density of nonzero elements (or equivalently a high proportion of zeros) are said to be *sparse*.

Sparsity is important since almost every operation involving matrix computations can be accelerated by only performing required operations on the nonzero elements. For example, an n by n dense matrix M leads to order of n^3 computations to produce the matrix M^2 , while it may require only order of n computations to produce the same matrix result with a sparse matrix. In other words, a successful sparse matrix implementation of the estimation problem may lead to computational work that is linear in n , while a dense matrix implementation could lead to work that rises with the cube of n .

78.5.3 Log-Determinant Calculations

There are a number of ways to attack the problem of calculating the log-determinant term in the log-likelihood function. These include closed-form solutions, eigenvalue approaches, Gaussian elimination approaches such as the Cholesky or LU decomposition, and approximations or bounds to the log-determinant. This section outlines these approaches and provides an empirical comparison in the last section.

Closed Form In some special cases, the log-determinant has a closed form. For example, many systems on regular grids have explicit closed-form solutions, or these can be easily extrapolated from a sequence of numerical determinants (LeSage and Pace 2009; Pace and LeSage 2009). Also, W based on just the single closest neighbor has a simple form of $n_{sp} \ln(1 - \lambda^2)$, where n_{sp} is the number of symmetric pairs of closest neighbors in W (Pace and Zou 2000). Finally, the log-determinant term vanishes when using a matrix exponential form of a spatial model, since the determinant of e^W equals $e^{\text{tr}(W)}$ which equals 1, and thus, the log-determinant equals 0 (LeSage and Pace 2007).

Eigenvalues Eigenvalues provide one of the most useful summaries of a matrix. For example, given the n by 1 vector of eigenvalues v from the spatial weight matrix W , the computation of the log-determinant in Eq. (78.51) and other quantities such as $\text{tr}(W^j)$ in Eq. (78.52) or $\text{tr}(WA^{-1})$ in Eq. (78.53) is simplified:

$$\ln |I_n - \lambda W| = \sum_{i=1}^n \ln(1 - \lambda v_i) \quad (78.51)$$

$$\text{tr}(W^j) = \sum_{i=1}^n v_i^j \quad (78.52)$$

$$\text{tr}(WA^{-1}) = \sum_{i=1}^n (1 - \lambda v_i)^{-1} v_i \quad (78.53)$$

Ord (1975) laid out the means of computing the spatial error model and the spatial lag of y model using eigenvalues and showed a useful similarity transformation that allows treating a row-stochastic (in linear algebra terms) W_{rs} (defined in Eqs. (78.54) and (78.55)) as a symmetric matrix in Eq. (78.56). Let R be an n by n diagonal matrix with the row sums of some underlying symmetric binary weight matrix B . In this case,

$$W_{rs} = R^{-1}B \quad (78.54)$$

$$\mathbf{1}_n = W_{rs}\mathbf{1}_n \quad (78.55)$$

$$W_{ss} = R^{-1/2}BR^{-1/2} \quad (78.56)$$

where the row-stochastic W_{rs} has the same real eigenvalues as the symmetric W_{ss} . In fact, both have a maximum eigenvalue of 1. This allows working with W_{ss} instead of W_{rs} when calculating the eigenvalues which increases the speed and typical accuracy of those calculations. Although there are some assertions in the literature that the eigenvalues of W_{rs} or W_{ss} cannot be calculated accurately for $n > 1,000$, this is rarely or ever true for W used in practice. Eigenvalue routines can encounter difficulties when the underlying eigenvalues are not distinct. If the underlying binary weight matrix B is symmetric, both W_{rs} or W_{ss} have distinct, real eigenvalues, and this will facilitate finding their eigenvalues. To give a specific example of the accuracy, we computed the eigenvalues for W_{ss} based on contiguity for 20,000 observations based on random locations. As a check, the sum of the eigenvalues should equal $\text{tr}(W_{ss}) = 0$, and these actually equaled 2.456×10^{-11} . In addition, we also compared the eigenvalue method to the Cholesky method (to be described shortly) and found for $\lambda = 0.9$, the difference in the two log-determinant values equaled 1.819×10^{-12} .

The main problem with eigenvalues is calculating these for large n . Most algorithms require dense matrices, which leads to storage issues as this requires working with n^2 elements. In the case of 20,000 observations, this uses 2.98 gigabytes per matrix, and in the case of the census tract example above with 62,226 observations, this would require 28.85 gigabytes per matrix. Moreover, most programs require at least twice the working space as the actual storage space needed. In addition, the calculation time rises with the cube of n , and so, this becomes another limiting factor. For example, it took 18.27 min to calculate the eigenvalues on the 20,000 by 20,000 matrix. The estimated time for a 62,226- by 62,226-sized matrix would be 9.17 h. Although this is becoming somewhat more feasible, many problems are still not feasible when using eigenvalues alone. Fortunately, there are alternative approaches that avoid these problems.

Gaussian Elimination Methods The quickest and most stable numerical method for finding log-determinants uses methods based on some form of Gaussian elimination. For symmetric W , this involves the Cholesky decomposition in Eq. (78.57) which reduces a symmetric, positive-definite matrix such as A into

the product of Cholesky triangular matrices U (the upper Cholesky triangle) in Eq. (78.57). The diagonal elements of U , U_{ii} for $i = 1 \dots n$, are termed the *pivots* and are all strictly positive. The sum of the log of the pivots equals the log-determinant of a triangular matrix. Because the Cholesky triangle is one version of the *square root* of A , the log-determinant of A is twice that of the log-determinant of U in Eq. (78.58):

$$I_n - \lambda W = U'U \quad (78.57)$$

$$\ln |I_n - \lambda W| = 2 \sum_{i=1}^n \ln(U_{ii}) \quad (78.58)$$

For nonsymmetric W , there is the related LU decomposition.

An advantage of the Cholesky or LU decompositions over the eigenvalue approach is that these methods take advantage of sparsity better than most eigenvalue methods. Especially, under certain orderings of the observations (George and Liu 1981; Pace and Barry 1997), the Cholesky triangles such as U stay relatively sparse.

Approximations A number of approximations to the log-determinant term have been proposed. For example, Martin (1993) proposed using the exact traces of the powers of W in the context of a power series as an approximation as in Eq. (78.59). However, the difficulty of computing the exact traces for dense W reduces the utility of the method:

$$|I_n - \lambda W| \approx \sum_{j=0}^m \lambda^j \text{tr}(W^j) \quad (78.59)$$

Barry and Pace (1999) built upon this approximation and used the additional approximation in Eq. (78.60), where M is a n by n matrix and u is a n by 1 vector composed of unit-independent normal deviates. This approximation rests on the properties of the unit normal *iid* distribution so that $E(u_i^2) = E(\chi_{1df}^2) = 1$ and $E(u_i u_j) = 0$ as shown in Eqs. (78.60)–(78.62):

$$\text{tr}(M) \approx u' M u \quad (78.60)$$

$$E(u_i M_{ii} u_i) = E(u_i^2 M_{ii}) = E(\chi_{1df}^2) M_{ii} = M_{ii} \quad (78.61)$$

$$E(u_i M_{ij} u_j) = 0 \quad (78.62)$$

With this method, they were able to approximate the log-determinant of a 1,000,000 by 1,000,000 matrix using a 133 MHz Pentium processor with 64 megabytes of memory. Note that $W^2 u$ is just $W(Wu)$ and $W^3 u$ is just $W(W^2 u)$ and so forth. Therefore, computing the moments just requires a sparse matrix–vector

Table 78.2 Times and accuracy across methods

n	Eigenvalues	Cholesky	MC	λ_c	λ_{mc}
1,000	0.167	0.087	0.073	0.873	0.872
2,500	2.431	0.058	0.076	0.877	0.876
5,000	18.371	0.141	0.141	0.876	0.876
10,000	136.372	0.443	0.261	0.864	0.864
15,000	448.853	0.670	0.329	0.869	0.869
20,000	1,096.045	0.967	0.475	0.870	0.870
50,000		2.584	1.323	0.869	0.869
100,000		5.628	2.415	0.868	0.868
1,000,000		132.649	22.747	0.871	0.871

operation. Also, given the estimated moments $\text{tr}(W^j)$, the overall log-determinant involves only a few calculations when recomputing $|I_n - \lambda W|$ for different values of λ . LeSage and Pace (2009, p. 99) also suggest using some of the lower ($j = 1, \dots, 4$ or $j = 1, \dots, 6$) exact moments as in Martin (1993) to improve accuracy. Some of these require little computational effort as $\text{tr}(W^2)$ for symmetric W is just the sum of all elements in W squared, which does not involve much work for sparse W . Zhang et al. (2007) improve on this algorithm, testing possible draws for quality by comparing the exact and approximate traces and retaining only the best draws.

Other approximations in the literature include approaches based on characteristic polynomials (Smirnov and Anselin 2001; Griffith 2004), Chebyshev polynomials (Pace and LeSage 2004), and extrapolation (Pace and LeSage 2009). Also, there are log-determinant bounds such as a simple quadratic inequality in Pace and LeSage (2002).

Comparisons Table 78.2 compares the time required to compute a vector of log-determinants for varying values of λ and shows the estimated dependence parameter $\tilde{\lambda}$ when using the Cholesky approach and the Monte Carlo log-determinant approximation based on four exact moments, 16 trials, and $m = 50$ iterations. As can be seen, using eigenvalues is feasible for problems involving up to 20,000 observations but becomes too demanding in both time and storage to proceed beyond that level. The Cholesky approach is feasible for all sample sizes, and the Monte Carlo log-determinant approximation works well for all sizes but only has a material advantage over the Cholesky approach for $n = 100,000$ or more. Using eigenvalues or the Cholesky approach yielded estimates of $\tilde{\lambda}$ that were equal to three decimal places, and the Monte Carlo log-determinant approximation led to the same $\tilde{\lambda}$ as the Cholesky for samples sizes of 2,500 or greater. In fact, rerunning the Monte Carlo log-determinant approximation with only two trials gave the same result as 16 trials but cut the time to 4.68 s from 22.75 s (on a 2.89 GHz Sandy Bridge CPU, Lenovo x220 laptop with 16 GB RAM).

Bivand (2010) compared the eigenvalue, Cholesky, Monte Carlo, and fifth degree Chebyshev log-determinant approaches and found they all performed well in terms of accuracy. In addition, he discussed and pointed out some errors in an earlier paper by Walde et al. (2008) regarding the various methods.

78.6 Conclusions

In terms of spatial econometrics for normally distributed (but dependent) disturbances, maximum likelihood methods have gone from being a challenging numerical problem limited to small sample sizes at the time of Ord (1975) to becoming routine and applicable to any sized data set. Part of this improvement over time arises from computational advances, but we should not overlook the role played by selecting approaches or techniques that are well suited to the case of spatial data when sparse connectivity structures are present. At this time, there are still challenges for more complicated likelihood problems such as those involving limited and dependent variables. It seems plausible that these currently challenging problems will become more routine over time as computational capacity expands and algorithms for specific problems improve.

Acknowledgments I would like to thank Mark Mclean, James LeSage, and Shuang Zhu for their very helpful comments.

References

- Anselin L (1988) Spatial econometrics: methods and models. Kluwer, Dordrecht
- Barry R, Pace RK (1999) LA Monte Carlo estimator of the log determinant of large sparse matrices, linear algebra and its applications 289(1–3):41–54
- Beron KJ, Vijverberg WPM (2004) Probit in a spatial context: a Monte Carlo analysis. In: Anselin L, Florax RJGM, Rey SJ (eds) Advances in spatial econometrics: methodology, tools and applications. Springer, Berlin/Heidelberg/New York, pp 169–195
- Bivand R (2010) Computing the Jacobian in spatial models: an applied survey (17 Aug 2010). NHH Department of Economics Discussion Paper No. 20/2010. Available at SSRN: <http://ssrn.com/abstract=1680467> or <http://dx.doi.org/10.2139/ssrn.1680467>
- Burridge P (2012) A research agenda on general-to-specific spatial model search. Invest Reg 21:71–90
- Chen J, Jennrich R (1996) The signed root deviance profile and confidence intervals in maximum likelihood analysis. J Am Stat Assoc 91(435):993–998
- Cramer JS (1986) Econometric applications of maximum likelihood methods. Cambridge University Press, Cambridge
- Davidson R, MacKinnon J (2004) Econometric theory and methods. Oxford University Press, New York
- George A, Liu J (1981) Computer solution of large sparse positive definite systems. Prentice-Hall, Englewood Cliffs
- Geweke J (1991) Efficient simulation from the multivariate normal and student-*t* distributions subject to linear constraints. In: Computer science and statistics: proceedings of the twenty-third symposium on the interface. American Statistical Association, Alexandria, pp 571–578
- Griffith D (1989) Advanced spatial statistics. Kluwer, Dordrecht
- Griffith D (2004) Faster maximum likelihood estimation of very large spatial autoregressive models: an extension of the Smirnov-Anselin result. J Stat Comput Simul 74(12):855–866
- Haining R (1990) Spatial data analysis in the social and environmental sciences. Cambridge University Press, Cambridge
- Hajivassiliou V, McFadden D (1990) The method of simulated scores for the estimation of LDV models with an application to external debt crises. Cowles Foundation Discussion Paper 967, Yale University

- Keane M (1994) A computationally practical simulation estimator for panel data. *Econometrica* 62(1):95–116
- LeSage JP, Pace RK (2007) A matrix exponential spatial specification. *J Econ* 140(1):190–214
- LeSage J, Pace RK (2009) Introduction to spatial econometrics. Taylor and Francis/CRC, Boca Raton
- Mardia KV, Marshall RJ (1984) Maximum likelihood estimation of models for residual covariance in spatial regression. *Biometrika* 71(1):135–146
- Martin RJ (1993) Approximations to the determinant term in Gaussian maximum likelihood estimation of some spatial models. *Commun Stat Theory Methods* 22(1):189–205
- Ord JK (1975) Estimation methods for models of spatial interaction. *J Am Stat Assoc* 70(1):120–126
- Pace RK, Barry RP (1997) Quick computation of spatial autoregressive estimators. *Geogr Anal* 29(3):232–246
- Pace RK, LeSage JP (2002) Semiparametric maximum likelihood estimates of spatial dependence. *Geogr Anal* 34(1):76–90
- Pace RK, LeSage JP (2004) Chebyshev approximation of log-determinants of spatial weight matrices. *Comput Stat Data Anal* 45(1):179–196
- Pace RK, LeSage J (2009) A sampling approach to estimating the log determinant used in spatial likelihood problems. *J Geogr Syst* 11(3):209–225
- Pace RK, LeSage J (2011) Fast simulated maximum likelihood estimation of the spatial probit model capable of handling large samples. Available at SSRN: <http://ssrn.com/abstract=1966039> or <http://dx.doi.org/10.2139/ssrn.1966039>
- Pace RK, Zhu S (2012) Separable spatial modelling of spillovers and dependence. *J Geogr Syst* 14(1):75–90
- Pace RK, Zou D (2000) Closed-form maximum likelihood estimates of nearest neighbor spatial dependence. *Geogr Anal* 32(2):154–172
- Phinikettos I, Gandy A (2011) Fast computation of high-dimensional multivariate normal probabilities. *Comput Stat Data Anal* 55(4):1521–1529
- Smirnov O, Anselin L (2001) Fast maximum likelihood estimation of very large spatial autoregressive models: a characteristic polynomial approach. *Comput Stat Data Anal* 35(8):301–319
- Walde J, Larch M, Tappeiner G (2008) Performance contest between MLE and GMM for huge spatial autoregressive models. *J Stat Comput Simul* 78(2):151–166
- Zhang Y, Leithead WE, Leithead DJ (2007) Approximate implementation of logarithm of matrix determinant in Gaussian processes. *J Stat Comput Simul* 77(4):329–348
- Zhu S, Pace RK Spatially interdependent mortgage decisions. *J Real Estate Fin Econ* (forthcoming)

Jeffrey A. Mills and Olivier Parent

Contents

79.1	Introduction	1571
79.2	Spatial Regression and Prior Modeling	1573
79.3	Bayesian Inference via MCMC	1574
79.3.1	A Brief Review of MCMC Theory	1575
79.3.2	Stationary Distributions and a Central Limit Theorem for MCMC	1576
79.4	MCMC Algorithms	1578
79.4.1	Gibbs Sampling	1578
79.4.2	Metropolis-Hastings (MH)	1579
79.4.3	Choice of Proposal Distribution	1580
79.5	Practical Considerations	1581
79.5.1	Setting Up and Monitoring MCMC Chains	1582
79.5.2	Other Tools and Post-Sampling Inference	1583
79.6	MCMC Inference for the SDM with Marginal Augmentation	1584
79.7	Spatiotemporal Model	1587
79.7.1	Empirical Application	1589
79.7.2	Estimation Results	1591
79.8	Conclusion	1593
	References	1594

Abstract

This chapter provides a survey of the recent literature on Bayesian inference methods in regional science. This discussion is presented in the context of the Spatial Durbin Model (SDM) with heteroskedasticity as a canonical example. The overall performance of different hierarchical models is analyzed. We extend the benchmark specification to the dynamic panel data model with spatial

J.A. Mills (✉) • O. Parent

Department of Economics, University of Cincinnati, Cincinnati, OH, USA
e-mail: jeffrey.mills@uc.edu; olivier.parent@uc.edu

dependence. An empirical illustration of the flexibility of the Bayesian approach is provided through the analysis of the role of knowledge production and spatiotemporal spillover effects using a space-time panel data set covering 49 US states over the period 1994–2005.

79.1 Introduction

Applied work in regional science is increasingly confronted with the task of analyzing data that are geographically referenced and temporally correlated, with many potential predictors. Up until the 1990s, virtually all of the empirical work in regional science employed frequentist statistical methods. The landmark work by Anselin (1988) reviews this literature and provides arguably the most comprehensive coverage of spatial econometrics in regional science.

In the early 1990s, the development of Markov Chain Monte Carlo (MCMC) methods revolutionized applications of the Bayesian approach to statistical inference. The revival of interest in the Bayesian approach has rapidly extended into spatial econometrics and geo-statistics. MCMC techniques, applied creatively, allow for the sophisticated modeling of large data sets with time dependence and cross-sectional correlation. Recent developments in Bayesian methods allow full Bayesian analysis of sophisticated multilevel models for complex geographically referenced data (Banerjee et al. 2004; LeSage and Pace 2009). This approach also offers full inference for non-Gaussian spatial data, spatiotemporal data, and, for the first time, solutions to problems of interpretation for models incorporating geographic and temporal dependence.

Analyzing a variety of panel data models, Chib (2008) underlines how the approach allows for the complex analysis of continuous, censored, count, and multinomial responses under weaker assumptions than required by previously developed methods. For instance, the Bayesian approach does not require the strict exogeneity assumption in the presence of endogenous covariates. Based on this panel setting, a growing number of studies examine spatial and temporal effects in multinomial or multivariate discrete response data. For example, Wang et al. (2012) develop a dynamic spatial ordered probit model and use it to analyze land development intensities. Discrete choice modeling with spatial dependence has been deeply analyzed using mainly the Bayesian approach (an extensive review can be found in LeSage and Pace 2009).

The development of new theoretical and empirical models in regional science to analyze, among other things, regional economic growth (Ertur and Koch 2007; LeSage and Fischer 2008), land use and conservation (Wang et al. 2012), industrial localization (Kakamu et al. 2012), geography of innovation (Autant-Bernard and LeSage 2011; Parent and LeSage 2008) highlights the flexibility of the Bayesian approach. There are additional problems that arise in the modeling process, such as model comparison and predictive performance, that have proven problematic in the

past, but can now also be addressed in a relatively straightforward manner using Bayesian inference and MCMC methods.

The rapid growth in availability of software incorporating MCMC methods has contributed to the dissemination and use of Bayesian methods in empirical work in regional science. A wide range of toolboxes contain all the standard procedures for empirical analysis. A comprehensive collection of routines can be found in one of the best known toolboxes for spatial data analysis, the spatial econometric toolbox of James LeSage (<http://spatial-econometrics.com/>). These routines are implemented within the MATLAB environment and contain the most advanced tools for spatial analysis and model interpretation. An increasingly attractive alternative is based on the development of statistical packages in the open source *R* environment (<http://r-project.org/>). An extensive collection of geo-statistics toolboxes are developed using Bayesian techniques. It is also worth mentioning the significant impact of open source software such as Winbugs. This has been used to make some significant contributions to empirical analysis in regional science.

This chapter presents recent econometric advances in the treatment of complex spatial and spatiotemporal data sets, and outlines a comprehensive approach to dealing with spatial and time effects from a Bayesian econometric perspective. The main objective is to illustrate how Bayesian techniques can help to understand a number of spatial theories and empirical models that have been developed for the practice of regional science and policy analysis. This discussion is presented in the context of the Spatial Durbin Model (SDM) as a canonical example.

The SDM is presented in the next section. Because the Bayesian method is inextricably tied to MCMC sampling, we provide a brief overview of MCMC methods in Sects. 79.3, 79.4, and 79.5. Section 79.6 then applies MCMC methods to the SDM to demonstrate some recent Bayesian research relevant for spatial econometric modeling, particularly with regard to problems of heteroskedasticity and spatial dependence in a panel data setting. The model is extended to include time dependence, and a substantive application of the methodology to regional growth models with interregional technological dependence is then provided in Sect. 79.7. Lastly, Sect. 79.8 summarizes and provides some concluding thoughts that relate to the future of Bayesian econometrics in regional science.

79.2 Spatial Regression and Prior Modeling

The Bayesian approach to spatial modeling relies extensively on the idea of a hierarchical prior which is used to model spatial dependence and heterogeneity. Suppose we have a cross-sectional sample of N independent observations y_i , $i = 1, \dots, N$ that are linearly related to a set of $N \times k$ explanatory variables X and are believed to be spatially correlated. As a benchmark, we will start with the Spatial Durbin Model, which can be motivated by concern over omitted variables or spatial heterogeneity (see LeSage and Pace 2009). This specification includes spatial lags of the explanatory variables as well as the dependent variable.

A representation of the Bayesian SDM is shown in Eq. (79.1):

$$\begin{aligned} y &= \rho W y + \iota_n \alpha + X\beta + WX\gamma + \varepsilon \\ \varepsilon &\sim N(0, \sigma_e^2 \Lambda^{-1}) \\ \Lambda &\equiv \text{diag}(1/\lambda_1, \dots, 1/\lambda_n) \\ \lambda_i &\sim \chi_v^2/v \end{aligned} \tag{79.1}$$

where W is a known $N \times N$ spatial weight matrix whose diagonal elements are zero, ι_n is a $1 \times N$ column vector of ones, and the strength of the spatial dependence is measured by the parameter ρ . The W matrix defines the structure of the dependence between (spatial) observational units. We also assume that W is normalized from a symmetric matrix, so that all eigenvalues are real and less than or equal to one. Different normalization methods can be used. For example, unlike the traditional row-normalization, the spectrally normalized matrix preserves the symmetry by dividing each element by the modulus of the largest eigenvalues (Barry and Pace 1999).

We add a normal-inverse gamma prior for β and σ_e , and we introduce a uniform prior distribution for the parameter ρ . Intuitively, if we were to simply treat Λ as N unrestricted parameters, a degrees-of-freedom problem would arise. Geweke (1993) proposes a set of N independent, identically distributed, chi-square distributions as prior information for the variance scalars λ_i ,

$$p(\Lambda) = \prod_{i=1}^n Ga\left(\lambda_i | \frac{v}{2}, \frac{v}{2}\right) \tag{79.2}$$

The parameter v represents the single parameter of the Gamma distribution equivalent to a chi-square distribution, allowing us to estimate the N variance scaling parameters λ_i by adding only a single parameter to the model. Geweke (1993) shows that this approach to modeling the disturbances is equivalent to a model that assumes a Student- t distribution for the errors. Another way to view this is that using a t distribution to deal with heteroskedasticity is equivalent to a scale mixture of normals when the mixing distribution is a Gamma distribution. That is, assuming that λ_i are independent $N(0, \sigma_e^2 \lambda_i^{-1})$ with prior for λ_i given in Eq. (79.2) is equivalent to the assumption that the error distribution is a weighted average of different normal distributions, each with a different variance. Additional flexibility in modeling heterogeneity can be achieved by introducing a prior hyperparameter for v that follows an exponential distribution governing the degrees of freedom that controls thickness of the tails in the Student- t error distribution (Geweke 1993).

79.3 Bayesian Inference via MCMC

As can be seen from Eq. (79.1), spatial models tend to have fairly high parameter dimensionality. This is because the minimal level of complexity needed to

adequately deal with variations in neighboring structure is rather high. As a result, analytical derivation of closed form expressions for Bayesian posterior distributions is not usually possible for these models. Fortunately, MCMC methods are a tailor-made solution to this problem as they provide approximations to posterior distributions in complex settings up to an arbitrary degree of numerical accuracy.

MCMC techniques allow simulation of a sample from any distribution by embedding it as a limiting distribution of a Markov chain, then simulating from the chain until it approaches equilibrium. This is essentially achieved by reverse engineering with the goal of finding a Markov chain algorithm that will ultimately converge upon the target distribution. Analogs of the law of large numbers and central limit theorems (see Sect. 79.3.2 below) exist for Markov chains that ensure that most of the simulated values from a chain can be used to provide information about the distribution of interest. The degree of accuracy can then be increased arbitrarily simply by increasing the simulated sample size.

A large theoretical literature now exists that sets out the conditions under which the MCMC chain converges to the target posterior. These conditions are surprisingly weak, though there is usually no way to guarantee that they hold in practice. However, a high degree of confidence that the Markov chain has converged can often be achieved, especially if care is taken to follow the suggestions in Geyer (2011) and employ the diagnostic tools discussed in Sect. 79.5 below.

In the last two decades, powerful MCMC techniques have been developed to obtain random draws from a very wide class of conditional distributions under remarkably general conditions. Even when the conditional distributions are too complex for Gibbs MCMC, Metropolis-Hastings (MH) algorithms can be employed to ensure that the appropriate limiting distribution is maintained by rejecting unwanted moves in a chain. We will assume the availability of algorithms to draw psuedo-random numbers from a variety of standard distributions. Methods for doing so have been thoroughly studied and are now widely available in most statistical software (see Gamerman and Lopes 2006, Chap. 1, for a good exposition).

79.3.1 A Brief Review of MCMC Theory

Monte Carlo methods originate from early work by Stanislaw Ulam and were used during World War II at Los Alamos in the development of the atomic bomb (Metropolis and Ulum 1949). Metropolis et al. (1953) was the pioneering paper on MCMC, but it was overlooked by statisticians, partly because it was published in a chemistry journal and partly because of the primitive level of computer technology available at the time, making computational methods prohibitively expensive for most statistical applications. Hastings (1970) generalized the Metropolis algorithm, but it was not until the late 1980s and early 1990s that widespread recognition of the practical importance of these algorithms occurred among statisticians. Geman and Geman (1984) developed the Gibbs sampler for use in image processing, and Tanner and Wong (1987) developed the data augmentation

approach and were arguably the first to recognize the potential for Bayesian MCMC inference. However, it was the classic expository paper by Gelfand and Smith (1990) that brought the Gibbs sampler to the attention of a wider audience. This led to the rapid development of a generic set of MCMC tools for Bayesian inference and subsequently revolutionized the field of statistics.

Most of the theoretical developments in MCMC were achieved in the 1990s. The research drive in MCMC methods over the last decade has shifted to developing more efficient computational tools. While advances in computer technology have continued to rapidly reduce the computational costs of simulation techniques, this has led to the analysis of more and more complex models. Researchers in MCMC methods continue to push the frontier of what is currently computationally feasible and there is need for a high level of computational efficiency in this environment. Many of the spatial models employed in empirical research, however, are not of such a high order of complexity, and can now be analyzed quickly and easily on any standard computer. Further, the level of complexity of the models that can be analyzed without being overly concerned with efficiency has grown dramatically in the last decade. In short, the last couple of decades have led to revolutionary changes in our ability to statistically analyze complex spatial models and problems.

79.3.2 Stationary Distributions and a Central Limit Theorem for MCMC

The goal of Bayesian computation is to obtain a sample of draws $\theta^{(t)}$, $t = 1, \dots, M$, from the posterior distribution of the unknown quantity θ , with a large enough sample that quantities of interest can be estimated with reasonable accuracy. MCMC simulation is a general method based on drawing values of θ from distributions that result in a sample from the target posterior distribution, $p(\theta|y)$. The sample is drawn sequentially, with the t th draw, $\theta^{(t)}$, depending only on the previous draw, $\theta^{(t-1)}$. This dependence on only the previous draw is the defining property of a Markov chain, which makes MCMC a practical application of Markov chain theory. Some understanding of the theory of Markov chains is thus helpful in practice, particularly in evaluating the performance and convergence of MCMC chains. This section provides a very brief review. Complete reviews can be found in many texts, including Gelman et al. 2004 and Gamerman and Lopes 2006.

A key requirement for the application of MCMC methods is the convergence of the chain to a stationary distribution. A distribution p is said to be a stationary distribution of a chain with transition probabilities $\pi = \pi(x, y)$ if $p = p\pi$. If the stationary distribution p exists and $\lim_{n \rightarrow \infty} p\pi^n = p$, then, independently of the initial distribution of the chain, p^n will approach p , as $n \rightarrow \infty$.

Ergodicity concerns ensuring that the chain will visit all possible values under the support of the distribution of interest (the stationary distribution) with nonzero probability. A chain is ergodic if it is aperiodic (so it cannot get stuck cycling in one subregion of the parameter space) and positive recurrent (which essentially means that as $n \rightarrow \infty$, the probability of visiting every possible state is nonzero).

For a Markov chain, $\theta = (\theta^{(1)}, \theta^{(2)}, \dots, \theta^{(n)})$, the ergodic average of a real-valued function of θ , $h(\theta)$ is the average $\bar{h}_n = (1/n) \sum_{t=1}^n h(\theta^{(t)})$.

If the chain is ergodic and $E_p[h(\theta)] < \infty$ for the unique limiting distribution p , then

$$\bar{h}_n \xrightarrow{a.s.} E_p[h(\theta)] \text{ as } n \rightarrow \infty \quad (79.3)$$

This result is a Markov chain equivalent of the Law of Large Numbers (see Gamerman and Lopes 2006, p. 125).

If a chain is uniformly (geometrically) ergodic and $h^2(\theta)(h^{2+\varepsilon}(\theta))$ is integrable with respect to p for some $\varepsilon > 0$, then we can obtain a Central Limit Theorem for Markov chains:

$$\sqrt{n} \frac{\bar{h}_n - E_p[h(\theta)]}{\tau} = \sqrt{n_{\text{eff}}} \frac{\bar{h}_n - E_p[h(\theta)]}{\sigma} \xrightarrow{d} N(0, 1) \text{ as } n \rightarrow \infty \quad (79.4)$$

where $\sigma^2 = \text{var}(h(\theta))$ is the variance of the limiting distribution p , $\tau^2 = \sigma^2(1 + 2 \sum_{k=1}^{\infty} \rho k)$ is the limiting sample variance of the estimate \bar{h}_n , and

$$n_{\text{eff}} = n \left/ \left(1 + 2 \sum_{k=1}^{\infty} \rho k \right) \right. \quad (79.5)$$

is the inefficiency factor due to autocorrelation in the Markov chain, indicated by $\rho_k = \text{cov}(h(\theta^{(t)}), h(\theta^{(t-k)}))/\sigma^2$. The inefficiency factor n_{eff} is used in practice to measure the “effective” random iid sample size of the MCMC chain by replacing the theoretical autocorrelations, ρ_k , with consistent sample estimates.

Equation (79.3) provides theoretical support for evaluating ergodic averages as estimates, and Eq. (79.4) supports evaluating approximate confidence intervals. Tierney (1994) provides proofs of ergodicity for the Markov chains in common use for MCMC simulation, so that the above results apply. See Gamerman and Lopes (2006) for further discussion.

One further point worth highlighting is the concept of a reversible Markov chain. A chain is said to be reversible if

$$p(x) = \pi(x, y) = p(y)\pi(y, x) \text{ for all } x, y \in S \quad (79.6)$$

where the state space S is the appropriate subset of \mathbb{R}^n representing the support of x, y .

Equation (79.6) is known as the “detailed balance equation” because it equates the rates of moves through states (so balanced) for every possible pair of states (hence detailed). This leads to the key result. If there is a distribution p satisfying the detailed balance equation (79.6), for an irreducible chain, then the chain is positive recurrent and reversible with stationary distribution p . Metropolis et al. (1953) showed that it is then *always* possible to construct a Markov chain with stationary distribution p by finding transition probabilities $\pi(x, y)$

satisfying Eq. (79.6). This provides an algorithm for constructing Markov chains that has weak requirements and so has wide applicability. The above results, in particular, convergence to the limiting distribution, the ergodic theorem, and the central limit theorem all hold for continuous state spaces with only minor technical modifications required (see Gamerman and Lopes 2006).

The above theory provides the means by which sampling from virtually any posterior distribution p can be achieved. The basic Metropolis algorithm is to set p as the limiting distribution of an ergodic Markov chain with transition kernel π . The various algorithms that build on this, in particular Gibbs sampling and Metropolis-Hastings (MH), are concerned with various methods of providing proposal distributions π to be sampled from.

79.4 MCMC Algorithms

The main workhorse MCMC method is Gibbs sampling, which is a special case of the MH algorithm that is very simple to use in practice. The Gibbs sampler requires knowledge of the full conditional distributions (up to an unknown constant) and so is not always usable, but simplifies the task and speeds up MCMC computations when it can be used. The MH algorithm does not require knowledge of the full conditionals and is often used in conjunction with the Gibbs sampler to obtain draws for the unknown parameters for which the full conditionals are not available.

79.4.1 Gibbs Sampling

The Gibbs sampler is an MCMC method that has wide applicability in spatial econometric modeling. Suppose we have a set of k parameter vectors, $\theta_1, \theta_2, \dots, \theta_k$, where each θ_i could be a scalar or a vector of parameter (to be drawn as a block). For example, in a linear regression model $y = X\beta + e$, $e \sim N(0, \sigma^2 I)$, it is convenient to separate the unknown parameters into two blocks, treating the regression coefficients as one ($1 \times k$) vector, so $\theta_1 = \beta$, and the variance separately as a scalar, $\theta_2 = \sigma^2$.

The Gibbs sampler can be used if we can sample from the full conditionals. The generic Gibbs sampler algorithm is to draw one value for each θ_i from its conditional distribution and cycle through these conditionals repeatedly. For each iteration, $t = 1, 2, \dots, M$, and arbitrary starting values $\theta_2^{(0)}, \theta_3^{(0)}, \dots, \theta_k^{(0)}$, the algorithm is,

- Draw $\theta_1^{(t)}$ from $p(\theta_1^{(t)} | \theta_2^{(t-1)}, \theta_3^{(t-1)}, \dots, \theta_k^{(t-1)}, y)$.
- Draw $\theta_2^{(t)}$ from $p(\theta_2^{(t)} | \theta_1^{(t)}, \theta_3^{(t-1)}, \dots, \theta_k^{(t-1)}, y)$.
- ⋮
- Draw $\theta_k^{(t)}$ from $p(\theta_k^{(t)} | \theta_1^{(t)}, \theta_2^{(t)}, \dots, \theta_{k-1}^{(t)}, y)$.

The above conditional distributions are the transition distributions of a Markov chain that converges (under very general conditions) to a unique stationary target

distribution that is the posterior distribution $p(\theta_i|y)$. In the linear regression example, we typically specify a normal prior distribution for β and an inverted-gamma prior for σ^2 . The Gibbs sampler for this model is then to cycle through the two conditionals, drawing $\beta^{(t)}$ from $N(\beta^{(t)}|\sigma^{2(t-1)}, y)$ and $\sigma^{2(t)}$ from $IG(\sigma^{2(t)}, |\beta^{(t)}, y)$.

The Gibbs sample for each parameter, $\theta_i^{(t)}$, $t = 1, 2, \dots, M$ then approximates a sample from the marginal posterior $p(\theta_i|y)$. This approximation can be made arbitrarily accurate by increasing the sample size, M . Given that it is now computationally inexpensive to obtain tens of thousands of draws on any standard computer for all but the most complex and highly dimensional models, Gibbs sampling is an easy way to draw posterior inferences concerning any unknown quantities in a model.

79.4.2 Metropolis-Hastings (MH)

MH algorithms are a general family of MCMC methods that use simulations from almost any arbitrary density π to actually generate draws from an equally arbitrary given target density p . Further, these algorithms allow for the dependence of π on the previous simulation, so the choice of π does not require a particularly elaborate construction a priori, but can take advantage of the local characteristics of the stationary distribution.

The use of a chain produced by an MCMC algorithm with stationary distribution p is fundamentally identical to the use of an iid sample from p in the sense that the ergodic theorem guarantees the (almost sure) convergence of the empirical average to the posterior expectation:

$$\frac{1}{M} \sum_{t=1}^M h(\theta^{(t)}|y) \xrightarrow{\text{a.s.}} E_p[h(\theta|y)]$$

A sequence $\theta^{(t)}$ produced by an MCMC algorithm can thus be employed just as an iid sample. An excellent introduction to Metropolis-Hastings algorithms is provided by Chib and Greenberg (1995).

79.4.2.1 The Metropolis Algorithm

The Metropolis et al. (1953) algorithm is a special case of the MH algorithm which draws from a transition distribution $\pi(\theta^{(t)}|\theta^{(t-1)})$ that must be symmetric, i.e., $\pi(\theta^{(t)}|\theta^{(t-1)}) = \pi(\theta^{(t-1)}|\theta^{(t)})$. This simplifies the algorithm in that the proposed transition distribution does not need to be evaluated at each accept-reject step since it does not appear in α (see below). Starting values, $\theta^{(0)}$, are often simply arbitrarily chosen to represent a draw from a preliminary crude approximate estimate of the posterior distribution, or are drawn from the prior distribution. Several runs of the algorithm using different starting values can be employed to diagnose convergence to the target posterior. Given starting values, for $t = 1, 2, \dots, M$, the algorithm is

- Draw $\theta^{(t)}$ from the transition distribution $\pi(\theta^{(t)}|\theta^{(t-1)})$.
- Calculate

$$\alpha = \frac{p(\theta^{(t)}|y)}{p(\theta^{(t-1)}|y)}$$

- Accept $\theta^{(t)}$ with probability = $\min(\alpha, 1)$, otherwise set $\theta^{(t)} = \theta^{(t-1)}$ (i.e., keep the previous draw).

This last step is accomplished by drawing a uniform random variate r in the $[0,1]$ interval and accepting $\theta^{(t)}$ if $\min(\alpha, 1) \geq r$.

The algorithm requires the ability to calculate the acceptance-rejection ratio α for all $(\theta^{(t)}, \theta^{(t-1)})$ and to draw $\theta^{(t)}$ from the proposal distribution $\pi(\theta^{(t)}|\theta^{(t-1)})$ for all θ and t . To prove that the sequence $\theta^{(t)} = t, 1, 2, \dots$ converges to a sample from the target distribution, we need (a) that the simulated sequence is a Markov chain with a unique stationary distribution and (b) that this stationary distribution equals the target posterior distribution. This holds if the Markov chain is irreducible, aperiodic, and nontransient. Except for trivial exceptions, the distribution is aperiodic and nontransient for a random walk on any proper distribution, and is irreducible if the random walk has a positive probability of eventually reaching any state from any other state (i.e., the transition distribution must be able to eventually visit all possible states with nonzero probability). The acceptance step and definition of α ensures, by construction, that the stationary distribution is the target posterior (see Gelman et al. 2004).

79.4.2.2 Metropolis-Hastings Algorithm

Hastings (1970) developed the MH algorithm as a generalization of the Metropolis algorithm such that the transition distribution is not required to be symmetric. In this case, the acceptance rule becomes

$$\alpha = \frac{p(\theta^{(t)}|y)/\pi(\theta^{(t)}|\theta^{(t-1)})}{p(\theta^{(t-1)}|y)/\pi(\theta^{(t-1)}|\theta^{(t)})}$$

Allowing asymmetric accept-reject rules can be useful in increasing the speed of convergence of the Markov chain. Proof of convergence to a unique stationary distribution is the same as for the Metropolis algorithm. That this stationary distribution is the target distribution follows from the definition of α (see Gelman et al. 2004). The Gibbs sampler can also be shown to be a special case of the MH algorithm with $\alpha = 1$ always, where the transition distribution is selected to be the conditional distribution $p(\theta^{(t)}|\theta^{(t-1)}|y)$.

79.4.3 Choice of Proposal Distribution

A good transition distribution is one for which, for any θ , it is easy to sample from $\pi(\theta^{(t)}|\theta^{(t-1)})$, it is easy to compute α , each accepted iteration moves a reasonable

distance in the parameter space (so that the Markov chain does not move too slowly), and the rejection rate is not too high (so that the chain does not remain in the same place too often). Note that only the ratios $\pi(\theta^{(t)}|\theta^{(t-1)})/\pi(\theta^{(t-1)}|\theta^{(t)})$ and $p(\theta^{(t)}|y)/p(\theta^{(t-1)}|y)$ are required, so we only need the kernels of these distributions.

While there are an infinite variety of possibilities, there are two main methods typically used for selecting the transition distribution. Random walk MH employs a transition distribution centered at the previous draw, so the draws follow a random walk over the support of the posterior. It is the most commonly used method because of its simplicity, its validity in most situations, and it does not require in-depth a priori knowledge of the transition distribution. The main alternative is the independent draw MH, which can be considerably more efficient, but requires a transition distribution that is a close approximation to the target distribution. The MH acceptance step is used to correct the approximation in the independent draw MH, with the goal being to accept as many draws as possible. If the posterior can be approximated fairly accurately with some confidence, then using the independent MH makes a lot of sense. Otherwise, the random walk MH tends to be the default choice.

The random walk MH with a normal transition kernel centered on the current draw, and with covariance matrix $= c^2 \hat{\Sigma}$, where $\hat{\Sigma}$ is an approximate estimate of the posterior covariance matrix, has transition matrix

$$\pi(\theta^{(t)}|\theta^{(t-1)}) \sim N(\theta^{(t-1)}, c^2 \hat{\Sigma})$$

The algorithm is then

- Start with $\theta^{(0)}$.
- Draw $\theta^{(t)} = \theta^{(t-1)} + \varepsilon$, $\varepsilon \sim N(0, c^2 \hat{\Sigma})$.
- Compute $\alpha = \min\{1, p(\theta^{(t)}|y)/p(\theta^{(t-1)}|y)\}$.
- With probability α , accept $\theta^{(t)}$, otherwise set $\theta^{(t)} = \theta^{(t-1)}$.
- Repeat as necessary.

The most efficient choice of the scale term for the normal random walk MH is $c \approx 2.4/\sqrt{k}$, where k is the dimension of θ (the number of parameters). This parameter, c , can be tuned by initial runs of the MH algorithm so that the acceptance rate is between 0.2 and 0.5, with the upper end appropriate in one dimension and the lower end for higher dimensions ($k > 5$), according to Gelman et al. (2004). While this algorithm can be improved in many ways, it has proved effective in many problems even with moderately large $k \lesssim 50$.

The independent draw MH takes the transition distribution to be independent of the current chain, so $\pi(\theta^{(t)}|\theta^{(t-1)}) = \pi(\theta^{(t)})$, and $\theta^{(t)}$ is drawn directly from this distribution, replacing the random walk step in the above algorithm. If $\pi(\theta^{(t)})$ is a good approximation to $p(\theta^{(t)}|y)$, then most draws will be accepted and we obtain a chain with almost no autocorrelation.

79.5 Practical Considerations

In practical application, both MCMC and Bayesian inference involve a number of choices concerning various parameters that must be selected a priori. The need to select prior distributions has, at least in the past, been a conceptual hurdle that slowed the widespread acceptance of Bayesian theory. With regard to MCMC, choice of burn-in sample size, tuning acceptance-rejection rate, length of MCMC chain needed, whether to use one chain or parallel chains, use every subsequent (accepted) draw or only keep every k th draw (and hence choose k), and appropriate choice and monitoring of convergence diagnostics represent only a partial list of the decisions the applied researcher has to make.

Fortunately, most of the anguish over these questions that was present in the 1990s has subsided as a combination of theoretical advances and practical experience provided reasonable answers. Bayesian inference and MCMC techniques have something of a parallel recent history in this regard. Development and extensive use of a widely accepted standard menu of relatively noninformative proper priors, coupled with demonstration of the robustness of posterior inference to reasonable variations in the parameters of these priors, along with many practical examples of their use, has essentially eliminated the controversy over the use of priors and hence Bayesian inference (see, e.g., Gelman et al. 2004). During the same period, appropriate procedures and choices for setting up, fine-tuning, and monitoring MCMC chains have become routine.

The two main practical issues that arise when using MCMC are as follows:

- (i) The early iterations can be misrepresentative of the target distribution since approximate convergence is likely to not have been reached yet; so inclusion of these early iterations will influence the posterior inference. We must therefore be sure to run the simulation algorithm for long enough to be confident that approximate convergence has been achieved and discard the early (burn-in) portion of the sample.
- (ii) The Markov chain can often be correlated. Inference from correlated draws is less precise than from the same number of independent draws because there is less new information in each correlated draw. Correlation in the draws can therefore make the sampling algorithm inefficient if a large number of draws is necessary to achieve a relatively small effective equivalent sample size of independent draws. To monitor this, we view the autocorrelation function (ACF) and calculate the effective sample size, Eq. (79.5).

We outline these procedures and give further references below. Geyer (2011) is an essential reference for anyone using MCMC methods in practice.

79.5.1 Setting Up and Monitoring MCMC Chains

Theoretically at least, many of the apparent problems that were of concern initially have turned out to be easily resolved. There is no theoretical justification for using any burn-in period, using parallel chains instead of just one chain, not using all

subsequent draws, or even for many of the convergence diagnostics originally developed. The short answer to all these issues is that one should simply run one chain for a longer time (number of iterations) to gain more confidence concerning convergence. Geyer (2011) argues that using a single longer chain is the best approach once variations in starting values have been explored. If long burn-in periods are required, or if the chains have very high autocorrelations, using a number of smaller chains may result in each not being long enough to be of any value. Where nonconvergence could be an issue (i.e., nonstandard problems), Geyer recommends at least one run of an MCMC chain overnight – “what better way for your computer to spend its time?” (Geyer 2011, p. 19).

The Gibbs sampler is the simplest of the MCMC algorithms and so is usually employed if sampling from the conditional posterior distributions is possible. If it is not possible to use the Gibbs sampler, the random walk Metropolis algorithm provides a relatively simple way to obtain an MCMC sample since we do not need to evaluate the transition distribution in the acceptance step. The computational power now available to the average user is such that obtaining MCMC sample sizes up to order 10^6 is already a fairly trivial task for many standard models. As a result, efficiency is no longer a real concern in many practical applications. In addition, a few easily implemented diagnostic tools have become standard, mainly:

- (a) Visual inspection of the chain itself (a simple time plot) to observe if the chain appears to have settled into a stationary path
- (b) Inspection of the ACF for the chain to check for excessive time dependence, requiring a larger number of draws (checking the effective sample size of independent draws by viewing the ACF for every k th draw)
- (c) Initially running the chain several times from a diverse set of starting values to check if the chain converges to the same stationary path each time
- (d) Tuning the acceptance rate for any MH steps to be somewhere between about 0.2 and 0.5
- (e) Calculation of numerical standard errors (NSEs) and an estimate of the effective sample size, n_{eff} , from Eq. (79.5)

A number of excellent monographs now exist that cover these issues in far more detail than is possible here. Of particular relevance for spatial modeling are Chib (2008) and, especially, LeSage and Pace (2009).

79.5.2 Other Tools and Post-Sampling Inference

When running an MCMC chain, the number of iterations should never be fixed in advance. Deciding on the length of an MCMC run is a sequential process where the MCMC chains are examined after pilot runs and new simulations (or new samplers) are chosen on the basis of these pilot runs. For many situations, an MCMC sample of 100 independent draws is sufficient for reasonable posterior summaries, so even with a fairly high degree of correlation in the chain, several thousand draws are generally more than sufficient for accurate posterior inference, provided we are

confident that the chain has converged (see Gelman et al. 2004). Further, we can compare sample standard errors with numerical standard errors to ensure the numerical accuracy is adequate, and run the chain for longer if it is not.

Once an MCMC sample is obtained, standard sample estimates of posterior moments and quantiles can be calculated for the unknown quantities directly, e.g., the posterior mean of any function, $h(\theta)$, of the unknown parameter θ , is estimated up to an arbitrary degree of numerical accuracy by

$$\bar{h}(\theta) = \sum_{t=1}^M h(\theta^{(t)})/M$$

The marginal posterior distribution can be examined by viewing histogram plots of the MCMC sample or fitting a smoothed kernel density estimate to the sample frequencies.

A widely used approach that reduces the variance of these estimators, especially useful for quantiles and tail area calculations, is known as Rao-Blackwellization, as it is derived from the application of the Rao-Blackwell Theorem. It can be shown that if the posterior conditional on some other parameter in the model, ϕ , can be evaluated using the MCMC samples for both θ and ϕ , the estimator

$$\bar{h}_\phi(\theta) = \sum_{t=1}^M \sum_{j=1}^M h(\theta^{(t)}) p(\theta^{(t)} | \phi^{(j)}, y) / M$$

dominates the unconditional sample estimator defined previously, $\bar{h}(\theta)$, in terms of variance (and squared error loss).

79.6 MCMC Inference for the SDM with Marginal Augmentation

For the Student- t SDM, as given by Eq. (79.1), the Gibbs sampler can be slow to converge because of posterior dependence among the variance parameters of σ_e^2 and Λ . Paradoxically, adding an additional parameter can improve the speed of convergence of the Markov chain simulation. This marginal augmentation or parameter expansion is a technique developed by Meng and Van Dyk (1999) to improve the rate of convergence of the MCMC algorithm. The idea is to reduce the correlation between draws via a working parameter that is not part of the original observed data model. Unlike conditional augmentation, where the working parameter is fixed at a specific value, marginal augmentation minimizes the augmented information by marginalizing over the working parameter. Note that not introducing a working parameter is, in fact, implicitly conditioning on a specific value. Avoiding this conditioning by modeling and integrating out that working parameter can increase the variability in the augmented data and thus reduces the augmented information. Data augmentation and parameter expansion methods dramatically increase the generality and applicability of this approach.

Focusing on the Student-t SDM defined in Eq. (79.1), convergence can be very slow between the homoskedastic variance σ_ε^2 and the heteroskedastic term Λ . If a posterior draw for σ_ε^2 is close to zero, then the draw for Λ will also be sampled with values near zero, and so on. Following Meng and Van Dyk's parameter expansion approach, we can reduce the correlation by adding a new working parameter whose only role is to allow the Gibbs sampler to move in more directions and thus improve the convergence. To accomplish this, we rewrite Eq. (79.1) as

$$\begin{aligned} y &= \rho Wg + \iota_n \alpha + X\beta + WX\gamma + \frac{\sigma_\varepsilon^2 z}{\sqrt{\Lambda}} \\ z &\sim N(0, I_N) \\ \Lambda &\equiv \text{diag}(\lambda_1, \dots, \lambda_n) \\ \lambda_i &\sim \chi_v^2/v \end{aligned} \tag{79.7}$$

The expanded model is

$$\begin{aligned} y &= \rho Wy + \iota_n \alpha + X\beta + WX\gamma + \frac{\sqrt{\omega} \sigma_\varepsilon^2 z}{\sqrt{q}} \\ z &\sim N(0, I_N) \\ q &\equiv \text{diag}(q_1, \dots, q_n) \\ q_i &\sim \omega \chi_v^2/v \end{aligned} \tag{79.8}$$

The parameter $\omega > 0$ can be viewed as an additional scale parameter. In this new specification, q plays the role of $\omega\Lambda$. Thus, introducing ω does not alter the model we are fitting.

Note that since $q_i = \omega\lambda_i$, then λ_i corresponds to q_i when $\omega = 1$. We expect marginal augmentation with a working prior independent of $\theta = (\beta, \rho, \sigma_\varepsilon^2)$ to improve the rate of convergence. We choose $p(\omega)$ to be the proper conjugate prior for ω in $p(Y, \lambda|\theta, \omega)$, namely, $\delta\chi_\gamma^{-2}$, where $\delta > 0$, $\gamma > 0$. As in Meng and van Dyk (1999), we use the standard improper prior $p(\beta, \log \sigma_\varepsilon^2) \propto 1$. Under this prior, Geweke (1993) shows that the posterior mean and standard deviation exist only if the prior $p(v)$ is null in the interval 0–4. We will assume the latter prior to be exponential $p(v) = \exp(v_0)$. The MCMC algorithm for this expanded model has the following steps:

- (a) $q_i|\beta, \sigma_\varepsilon^2, \rho, \omega, Y \sim \frac{\omega}{(y_i - \rho \sum_j w_{ij}y_j - x_i\beta)^2 / \sigma_\varepsilon^2 + v} \chi_{v+1}^2$ independently for $i = 1, \dots, n$.
- (b) $\omega|Y, q \sim \frac{\delta+v \sum_{i=1}^n q_i}{\chi_{v+n}^2}$.
- (c) $\beta|\sigma_\varepsilon^2, Y, q, \rho, \omega \sim N(c, T)$, where $c = (X'qX)^{-1}X'q(I_n - \rho W)y$ and $T = \omega \sigma_\varepsilon^2 (X'qX)^{-1}$.
- (d) $\sigma_\varepsilon^2|Y, q, \omega \sim \frac{\sum_{i=1}^n q_i(y_i - \rho \sum_j w_{ij}y_j - x_i\beta)^2}{\omega \chi_{n+1}^2}$.

$$(e) \rho|\beta, \sigma_\varepsilon, q \propto |I_n - \rho W| \exp\left\{-\frac{1}{2\sigma_\varepsilon^2} e'qe\right\}, \text{ where } (I_n - \rho W)y - X\beta.$$

$$(f) v|y, q \propto \left(\frac{v}{2}\right)^{\frac{Nv}{2}} \Gamma\left(\frac{v}{2}\right)^{-N} \exp(\eta v), \text{ with } \eta = \frac{1}{v_0} + \frac{1}{2} \sum_{i=1}^N [\ln(q_i^{-1} + q_i)].$$

The first four steps involve Gibbs sampling, but the last two posterior densities are non-standard, and a Metropolis-Hastings step is implemented. A random walk MH algorithm with a normal increment random variable is therefore used for these steps, as described in [Sect. 79.4.2](#).

Given these conditional distributions, we can implement a data augmentation algorithm with marginal augmentation. At iteration $(t + 1)$, we draw $q^{(t+1)}$ from the conditional with marginal augmentation,

$$p(q|\beta, \sigma_\varepsilon^2, \rho, Y) = \int p(q|\mu, \sigma_\varepsilon^2, Y, \omega)p(\omega)d\omega \quad (79.9)$$

The implementation of this marginal augmentation is performed using the following scheme:

- (a) Step 1: Draw ω from its prior $p(\omega)$ and then q from $p(q|\beta, \sigma_\varepsilon^2, Y, \omega)$.
- (b) Step 2: Given q , $(\beta, \sigma_\varepsilon, \rho)$ is generated from $p(\beta, \sigma_\varepsilon, \rho|Y, q) = \int p(\beta, \sigma_\varepsilon, \rho|Y, q)p(\omega|Y, q)d\omega$ by first drawing ω from the posterior $p(\omega|Y, q)$, then drawing $\beta, \sigma_\varepsilon$, and ρ given ω their posterior distributions.

As a comparison, the conditional augmentation approach would fix $\omega = 1$ and ignore its posterior distribution.

A Monte Carlo experiment was conducted to evaluate the performance of the above sampling method and compare the conditional versus the marginal augmented methods. The data generating process is shown in Eq. (79.10).

$$\begin{aligned} y &= \rho Wy + \iota_n \alpha + X\beta + WX\gamma + \frac{\sigma_\varepsilon^2 z}{\sqrt{\Lambda}} \\ z &\sim N(0, I_N) \\ \Lambda &= \text{diag}(\lambda_1, \dots, \lambda_n) \\ \lambda_i &\sim \chi_v^2/v \end{aligned} \quad (79.10)$$

The spatial weight matrix W was generated using random points in conjunction with the MATLAB Delaunay routine to produce a symmetric contiguity weight matrix that is then row-normalized (see Chap. 4 in LeSage and Pace [2009](#)). Explanatory variables x_{it} were generated from zero mean independent normal distributions with a variance of four ($N(0, 4)$). A discussion about the impact of the choice of the hyperparameters δ and γ for the prior distribution of ω can be found in Meng and van Dyk [\(1999\)](#). We set $\gamma = 2$ and $\delta = 0.001$. [Table 79.1](#) presents posterior means, standard deviations, and numerical standard error (NSE) measures of the accuracy associated with estimates based on marginal versus conditional data augmentation steps.

Table 79.1 Monte Carlo simulations for $n = 500$ and 1,000 iterations

Parameter	True value	Marginal DA, $\omega \sim \delta \chi_v^{-2}$			Conditional DA $\omega = 1$		
		Mean	S.D.	NSE	Mean	S.D.	NSE
ρ	0.7	0.6957	0.0086	0.6785	0.6977	0.0046	0.6882
β	1	1.0011	0.0096	0.9820	1.0020	0.0100	0.9820
γ	1	0.9722	0.0266	0.9820	1.0492	0.0352	0.9838
σ_e^2	1	0.9979	0.7109	0.3096	0.9434	0.7515	0.2782
v	4	4.0224	0.0111	4.0015	4.0456	0.0423	4.0015

The marginal augmentation is less sensitive to the choice of starting value and decreases the numerical standard errors. Conditioning on the working parameter $\omega = 1$, reduces the speed of convergence of the chain. The parameter of σ_e^2 clearly needs more iterations to be completely independent from the initial iterations under the conditional marginal data augmentation.

79.7 Spatiotemporal Model

In this section, the SDM with heteroskedasticity is extended into a dynamic panel data model that accommodates spatial dependence. A variety of models that control for serial correlation and spatial dependence across locations have been explored (see Lee and Yu 2010; for a complete review). Yu et al. (2008) analyze a specification that allows for both time and spatial dependence as well as a cross-product term reflecting spatial dependence at a one-period time lag. This last term can be interpreted as the spatial diffusion that takes place over time. Parent and LeSage (2012) extend this approach, introducing a space-time filter that can be applied to the dependent variable or the error term. This filter implies a constraint on the mixing term that reflects spatial diffusion or space-time covariance. Parent and LeSage (2012) show that this constraint allows for a number of simplifications in the Bayesian MCMC estimation scheme.

The space-time filter is applied to the following panel data model with random effects and heteroskedastic disturbances across locations:

$$\begin{aligned}
 y_t &= \iota_N \alpha + x_t \beta + W x_t \gamma + \eta_t \quad t = 0, \dots, T \\
 \eta_t &= \mu + \frac{\sigma_e^2 z_t}{\sqrt{\Lambda}} \\
 z_t &\sim N(0, I_N) \\
 \Lambda &\equiv \text{diag}(\lambda_1, \dots, \lambda_n) \\
 \lambda_i &\sim \chi_v^2 / v
 \end{aligned} \tag{79.11}$$

where $y_t = (y_{1t}, \dots, y_{Nt})'$ is the $N \times 1$ vector of observations for the t th time period, α is the intercept, ι_N is an $N \times 1$ column vector of ones, x_t denotes the $N \times k$ matrix of non-stochastic regressors and μ is an $N \times 1$ column vector of

random effects, with $\mu_i \sim N(0, \sigma_\mu^2)$. The random error terms $\varepsilon_t = \sigma_\varepsilon^2 z_t / \sqrt{\Lambda}$ are assumed to be independent and identically distributed with zero mean and a variance $\sigma_\varepsilon^2 \Lambda^{-1}$. We make the traditional assumption that μ is uncorrelated with ε_t , and Λ represents the heteroskedastic covariance matrix.

To define the time filter, let C be the Prais-Winsten transformation, where ϕ is the autoregressive time dependence parameter. This filter is defined as:

$$C = \begin{pmatrix} \psi & 0 & \dots & 0 \\ -\phi & 1 & \dots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \dots & -\phi & 1 \end{pmatrix} \quad (79.12)$$

Specification of ψ , the (1,1) element in C , depends on whether the first period is modeled or assumed to be known. For simplicity, we will assume the first observations to be known. The space filter is defined as a nonsingular matrix $B = (I_N - \rho W)$, where ρ is a scalar spatial dependence parameter and W is the known $N \times N$ spatial weight matrix as above. The proposed space-time filter then corresponds to the Kronecker product of the matrices C and B ,

$$C \otimes B = I_{N,T+1} - \rho I_{T+1} \otimes W - \phi L \otimes I_N + (\rho \times \phi) L \otimes W \quad (79.13)$$

where L is a $(T + 1) \times (T + 1)$ matrix based on the time-lag operator. This filter implies a restriction that the parameter associated with spatial effects from the previous period ($L \otimes W$) is equal to $-\rho \times \phi$. Parent and LeSage (2012) show that applying this space-time filter to the dependent variable greatly simplifies the estimation procedure when an optimal predictor is used to model the first observation. They also advocate that imposing this constraint would simplify the interpretation of the marginal effects.

We decide to ignore the issues pertaining to prediction of the first period cross section values, and apply the filter to the dependent variable resulting in a model specification

$$\begin{aligned} (C \otimes B)y &= I_{N,T+1}\alpha + x\beta + (I_{T+1} \otimes W)x\gamma + \eta \\ \eta &\sim N(0, \tilde{\Omega}) \end{aligned} \quad (79.14)$$

where $y = (y'_0, \dots, y'_T)'$, $x = (x'_0, \dots, x'_T)'$ and

$$\tilde{\Omega} = \sigma_\mu^2 (J_{T+1} \otimes I_N) + \sigma_\varepsilon^2 I_{T+1} \otimes \Lambda_N^{-1} \quad (79.15)$$

with $J_{T+1} = I_{T+1} l'_{T+1}$.

A number of studies have treated the parameter $\rho \times \phi$ associated with the cross-product term in different ways. Anselin (1988) proposed a related “time-space dynamic model” specification explored by Yu et al. (2008) who relaxed the implied

constraint $\theta = -\rho \times \phi$ and estimated an unrestricted parameter θ . We will start with this general specification and show that the constraint implied by the space-time filter is relevant and makes the model easy to interpret. Since we ignore the first period, the general panel data model specification with random effects for $t = 1, \dots, T$, is given by

$$\begin{aligned} y_t &= \phi y_{t-1} + \rho W y_t + q W y_{t-1} + \iota_N \alpha + x_t \beta + W x_t \gamma + \eta_t \\ \eta_t &= \mu + \varepsilon_t \end{aligned} \quad (79.16)$$

One advantage of the Bayesian MCMC scheme we propose for this specification is that it does not require integration over the random effects that appear in the likelihood. However, integration over these parameters can reduce serial dependence in the MCMC samples of parameters drawn. A formal expression of posterior distributions for this specification can be found in Parent and LeSage (2012). The only difference relies on the heteroskedastic term λ_i that is generated from the following chi-square distribution:

$$\lambda_i | \beta, \sigma_e^2, \rho, \omega, Y \sim \frac{\chi_{v+T}^2}{\sigma_e^{-2} e'_i e_i + v} \quad i = 1, \dots, n \quad (79.17)$$

where

$e_i = y_i - \phi y_{i,-1} - \rho \sum_{j=1}^N w_{i,j} y_j - \theta \sum_{j=1}^N w_{i,j} y_{j,-1} - \alpha \iota_T - x_i \beta - \sum_{j=1}^N w_{i,j} x_j \gamma$,
 $y_i = (y_i, 1, \dots, y_i, T)'$ and $y_i, -1 = (y_i, 0, \dots, y_i, T-1)'$. We also set the hyperparameter $v = 4$, consistent with our prior belief in heteroskedasticity.

79.7.1 Empirical Application

In this empirical illustration, we model and estimate the presence of interregional technological dependence. We rely on a simple model of semi-endogenous growth developed by Jones (2002). This empirical analysis shows that implementing spatial and time dependence conveys important information regarding to what extent innovative activities spill over to neighboring states.

Based on the model described by Jones (2002), we propose a dynamic specification where the stock of knowledge in the neighboring regions has spillover effects on the growth rate of ideas in region i :

$$\frac{\dot{A}_i(t)}{A_i(t)} = \delta L_i(t)^\lambda A_i(t)^{\gamma-1} \prod_{j \neq i} A_j(t)^{\psi w_{ij}} \quad (79.18)$$

According to Eq. (79.18), the number of new ideas produced at any point in time is driven by the number of researchers and the existing stock of ideas in region i as well as in neighboring regions. The parameter λ represents the effect of research on new ideas and allows for the possibility of duplication. For now, we assume $|\lambda| < 1$ and $|\psi| < 1$, but stability conditions are discussed in more detail later. Based on

Parent (2012), we define the connectivity structure W using a measure of both geographical as well as technological proximity. The spatial weight scheme is based on the concept of five nearest neighbors where these five neighbors will receive varying weights based on the measure of technological proximity.

Parent (2012) shows that using the log-linearization of Eq. (79.18) around the vector of steady state growth rates g_A , where $A(t)$ and $L(t)$ are growing at constant rates, corresponds to

$$\frac{\dot{A}_i(t)}{A_i(t)} = g_A(1 - \log(g_A/\delta)) + g_A \left[\lambda \log(L_i(t)) - (1 - \gamma) \log(A_i(t)) + \sum_{j \neq i} \psi w_{ij} \log(A_j(t)) \right] \quad (79.19)$$

and we can rewrite Eq. (79.19) as

$$\log(A_i(t+1)) = \phi \log(A_i(t)) + \sum_{j \neq i} \theta w_{ij} \log(A_j(t)) + \alpha + \beta \log(L_i(t)) \quad (79.20)$$

where $\phi = -g_A(1 - \gamma) + 1$, $\theta = g_A\psi$, $\alpha = g_A(1 - \log(g_A/\delta))$ and $\beta = g_A\lambda$.

The parameter θ captures the impact of accessible external ideas on regional innovative activities also called interregional knowledge spillovers. We can add to the econometric specification problems noted by Jones' (2002) omitted variables bias that would arise from excluding $\sum_{j \neq i} \theta w_{ij} \log(A_j(t))$ from the model by assuming $\psi = 0$, leading to $\theta = g_A\psi = 0$.

We extend the theoretical framework Eq. (79.20), where the diffusion process is similar to an autoregressive model where spatial interaction occurs with a lag of one period. We introduce the traditional simultaneous spatial lags used in cross-sectional models from the spatial econometrics literature, where the right-hand-side variable takes the form $\sum_{j \neq i} \rho w_{ij} \log(A_j(t+1))$.

The stock of patents per capita for state i at the period t results in the following regression:

$$\begin{aligned} \log(A_{it}) &= \alpha + \sum_{j \neq i} \rho w_{ij} \log(A_{j,t}) + \phi \log(A_{i,t-1}) + \sum_{j \neq i} \theta w_{ij} \log(A_{j,t-1}) \\ &\quad + \beta \log(L_{i,t-1}) + \gamma W \log(L_{i,t-1}) + \eta_{it} \end{aligned} \quad (79.21)$$

$$\eta_{it} = \mu_i + \varepsilon_{it}.$$

Externalities generated by one region are allowed to influence neighboring regions within the same (annual) time period (the spatial effect), the same region in subsequent periods (the time effect), as well as neighboring regions in subsequent periods (the space-time diffusion effect). This space-time dynamic allows us to compare the relative importance of contemporaneous spatial dependence with time dependence and spatial interaction from the previous periods.

To estimate this model, we must measure the stock of ideas. Observable measures of new ideas at a regional or international level are never perfect. We organize the analysis by focusing on the observed number of US domestic patents, a useful indicator of the state level of realized innovation for a given period. We estimate the knowledge production function using a dataset on patenting activity and its determinants covering the period 1994–2005 and 49 states. (The District of Columbia is treated as a state and the states of Alaska and Hawaii are omitted.) The data include patents granted per capita for each state in each year along with measures of the factor inputs in the production function for ideas/knowledge.

Skilled labor $L_i(t)$ for each state i at time period t is measured using two explanatory variables: $doc_i(t)$, the number of doctoral recipients, and $expRD_i(t)$, total research and development expenditures as a percentage of gross state product. Total R&D expenditures are calculated by adding all sources of funds: industry, public and private nonprofit institutes, and universities.

79.7.2 Estimation Results

Estimation results are presented in [Table 79.2](#), based on a sample of 50,000 draws collected after a burn-in period of 10,000 draws. In the following discussion of the parameter estimates we relied on 5 and 95 percentage points of the highest posterior density intervals (HPDI) to draw inferences regarding whether the posterior means were different from zero.

As explained in LeSage and Pace ([2009](#)), low levels of spatial dependence between neighboring regions can over time lead to a significant amount of inter-connectivity between regions in the long-run knowledge production process. Ignoring the low levels of observed spatial dependence will have dramatic impacts on the long-run estimates and inference regarding the regional knowledge production and diffusion process.

Traditionally, a positive effect of spatial dependence is interpreted as local spillover effects related to the presence of knowledge stocks in neighboring regions. Parent and LeSage ([2008](#)) make the point (in the context of European regions) that positive spatial dependence of this type may arise when regions possess the ability to absorb and to adopt new technologies of their neighbors. Further, R&D activities can increase the incidence of technology diffusion by enhancing a region's absorptive capacity. Positive spatial dependence found here using the space-time model leads to an inference that R&D expenditures will directly increase the level of innovation occurring in a region over time.

In fact, as explained by Debarsy, Ertur, and LeSage ([2012](#)), a change to explanatory variable r at time t will have direct and indirect impacts on the own- and other-region-dependent variable values at time t , as well as impacts on both own and other regions in future time periods. This diffusion over space as time passes arises when the model includes nonzero time dependence captured by the parameter ϕ .

Turning to the restriction implied by the space-time filter $\theta = -\rho \times \phi$, estimation results presented in [Table 79.2](#) reveal that this restriction is consistent with the

Table 79.2 Estimation results

Parameter	Post. mean	S.D.	Lower 0.05	Upper 0.95
Constant	0.2949	0.1403	0.0949	0.5444
doc	0.0028	0.0130	-0.0161	0.0274
expRD	0.0602	0.0256	0.0214	0.1046
W doc	-0.0260	0.0168	-0.0587	0.0006
W expRD	-0.0354	0.0268	-0.0825	0.0040
ρ	0.4157	0.0353	0.3607	0.4807
ϕ	0.9152	0.0632	0.7320	0.9657
θ	-0.3798	0.0413	-0.4477	-0.2999
σ_{μ}^{-2}	0.0102	0.0113	0.0012	0.0354
$\sigma_{\varepsilon}^{-2}$	0.0145	0.0013	0.0125	0.0168

data for both specifications. The partial derivatives for this situation are shown in Eq. (79.22) for the case where we change the explanatory variable $x_t^{(r)}$ at time period 1, and measure the impacts at a one- through t -period horizon. Since the estimation results confirm the time-separability constraint $\theta = -\rho \times \phi$, the partial derivative can be rewritten as

$$\frac{\partial y_t}{\partial x_1^{(r)}} = (\phi^{t-1} + \phi^{t-2} + \dots + \phi + 1)B^{-1}(I_N\beta_r + W\gamma_r) \quad (79.22)$$

This greatly simplifies interpretability of the dynamic responses for any number of time periods. Given estimates for the parameters β_r , ρ , and ϕ , we can easily calculate dynamic responses for any number of time periods. In fact, the diffusion over time and space takes the form of time discounting based on the time dependence parameter ϕ of the contemporaneous spatial effects captured by the $N \times N$ matrix B^{-1} .

Table 79.2 shows scalar summary measures of the effects estimates for spatial dependence ($\rho = 0.42$) that is relatively weaker than time dependence $\phi = 0.92$, which leads to larger time and space-time diffusion effects relative to the spatial effects. Based on the stationary conditions defined by Parent and LeSage (2012), the process is stationary since $\phi + \rho + \theta < 1$ and $\phi - (\rho - \theta) > -1$.

Table 79.3 reports cumulative spatial effects decomposed into direct, indirect, and total effects. The direct effects correspond to own-partial derivatives that measure the impact on region i from changes in the explanatory variable value of region i . However, these include some feedback impacts discussed in LeSage and Pace (2009), since changes in region i influence the neighbors and region i is in turn influenced by its neighbors. The indirect effects are cumulated over neighboring spatial regions and correspond to the cross-partial derivatives, and the final column shows the total effects which is the sum of the direct and indirect effects. In our model, the spatial effects are separable from the time effects, and these do not change over time since the spatial configuration of the regions remains the same and

Table 79.3 Scalar summary estimates of the R&D effects

	Lower 0.05	Median expRD	Upper 0.95
Spatial effects			
Direct	0.0225	0.0632	0.1099
Indirect	0.0141	0.0398	0.0692
Total effects	0.0366	0.1030	0.1790
Cumulative effects			
Direct	0.1812	0.5096	0.8855
Indirect	0.1140	0.3208	0.5574
Total effects	0.2952	0.8304	1.4429

we restrict the spatial dependence parameter to be fixed over all time periods. The differences between the cumulative total effects and the spatial effects reflect the importance of the time effects. In the case of R&D expenditures, we see a 0.5096 direct cumulative effect value and a direct spatial effect of 0.0632, so the difference of 0.4464 represents cumulative direct time effects (which we calculated over a 14-year horizon). In comparison with the coefficient estimate of 0.0602 from Table 79.2 for this variable, the direct effects estimate reported in Table 79.3 includes a feedback loop that arises in our space-time dynamic panel model.

Consistent with the ideas-based growth literature, the results suggest that the level of innovation is positively influenced by the level of effort devoted to the ideas sector. Expenditures on R&D have a more permanent impact on the growth process if a highly skilled labor force eases the adoption of new technologies. Of course, this is consistent with the observation that regions with advanced levels of technology often have strong links with education, especially at the doctoral level. Thus, more education should lead to higher rates of technological progress via improvements in labor force quality. However for both models, the effect of the variable *LDoc* is not statistically significant.

As shown in Parent (2012), these results confirm that interactions between regions are spatially limited and localized spillovers effects can lead to regional clusters with persistently different levels of innovative activity.

79.8 Conclusion

This chapter shows how the Bayesian approach provides a complete inferential toolkit for a variety of cross-sectional and panel data spatial models. Bayesian methods have recently produced some remarkably efficient solutions to complex inference problems. The approach is based on a combination of hierarchical prior modeling and MCMC simulation methods. Interestingly, this approach is able to tackle estimation and model interpretation in situations that are quite challenging by other means.

Marginal data augmentation improves the convergence properties of the MCMC sampler. This method expands the parameter space with a working parameter that is only identifiable given the augmented data. Placing a prior distribution directly on the identifiable parameters results in enormous computational gain. This prior specification can make the model easier to estimate and interpret in many complex cases like multivariate and multinomial discrete choice models.

While this chapter is necessarily too brief to provide a self-contained guide, hopefully it sheds enough light on the main conceptual issues to demonstrate that using Bayesian MCMC inferential tools allows for broad generality in model specification, and is relatively simple to use in practice. The growth of Bayesian MCMC spatial econometric methods continues at a rapid pace as the Bayesian approach becomes more widely understood and as software and computing power become more readily available.

References

- Anselin L (1988) Spatial econometrics: methods and models. Kluwer, Boston
- Autant-Bernard C, LeSage JP (2011) Quantifying knowledge spillovers using spatial econometric models. *J Reg Sci* 51(3):471–496
- Banerjee S, Carlin B, Gelfand A (2004) Hierarchical modeling and analysis for spatial data. Chapman & Hall, Boca Raton
- Barry R, Pace RK (1999) Monte Carlo estimates of the log determinant of large sparse matrices. *Linear Algebra Appl* 289(1–3):41–54
- Chib S (2008) Panel data modeling and inference: a Bayesian primer. In: Matyas L, Sevestre P (eds) *The econometrics of panel data*. Springer, Berlin/Heidelberg, pp 479–515
- Chib S, Greenberg E (1995) Understanding the Metropolis-Hastings algorithm. *Am Stat* 49(4):327–335
- Debarsy N, Ertur C, LeSage JP (2012) Interpreting dynamic space-time panel data models. *Stat Methodol* 9(1–2):158–171
- Ertur C, Koch W (2007) The role of human capital and technological interdependence in growth and convergence processes: international evidence. *J Appl Econom* 22(6):1033–1062
- Gamerman D, Lopes HF (2006) *Markov chain Monte Carlo*. Chapman & Hall
- Gelfand AE, Smith AFM (1990) Sampling-based approaches to calculating marginal densities. *J Am Stat Assoc* 85(410):98–409
- Gelman A, Carlin JB, Stern HS, Rubin DB (2004) *Bayesian data analysis*, 2nd edn. Chapman & Hall, London
- Geman S, Geman D (1984) Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Trans Pattern Anal Mach Intell* 6(6):721–741
- Geweke J (1993) Bayesian treatment of the independent Student-t linear model. *J Appl Econom* 8(1):519–540
- Geyer C (2011) Introduction to Markov chain Monte Carlo. In: Brooks SP, Gelman A, Jones G, Meng X-L (eds) *Handbook of Markov chain Monte Carlo*. Chapman and Hall/CRC Press, Boca Raton
- Hastings WK (1970) Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* 57(1):97–109
- Jones CI (2002) Sources of U.S. economic growth in a world of ideas. *Am Econ Rev* 92(1):220–239
- Kakamu KW, Polasek W, Wago H (2012) Production technology and agglomeration for Japanese prefectures during 1991–2000. *Paper Reg Sci* 91(1):29–41

- Lee LF, Yu J (2010) Some recent developments in spatial panel data models. *Reg Sci Urban Econ* 40(5):255–271
- LeSage JP, Fischer MM (2008) Spatial growth regressions: model specification, estimation and interpretation. *Spatial Econ Anal* 3(3):275–304
- LeSage JP, Pace RK (2009) An introduction to spatial econometrics. CRC Press, Boca Raton
- Meng XL, van Dyk DA (1999) Seeking efficient data augmentation schemes via conditional and marginal augmentation. *Biometrika* 86(2):301–320
- Metropolis N, Ulam S (1949) The Monte Carlo method. *J Am Stat Assoc* 44(247):335–341
- Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH, Teller E (1953) Equation of state calculations by fast computing machine. *J Chem Phys* 21:1087–1092
- Parent O (2012) A space-time analysis of knowledge production. *J Geogr Syst* 14(1):49–73
- Parent O, LeSage JP (2008) Using the variance structure of the conditional autoregressive spatial specification to model knowledge spillovers. *J Appl Econom* 23(2):235–256
- Parent O, LeSage JP (2012) Spatial dynamic panel data models with random effects. *Reg Sci Urban Econ* 42(4):727–738
- Tanner MA, Wong W (1987) The calculation of posterior distributions by data augmentation (with discussion). *J Am Stat Assoc* 82(398):528–550
- Tierney L (1994) Markov chains for exploring posterior distributions (with discussion). *Ann Stat* 22(4):1701–1762
- Wang X, Kockelman K, Lemp J (2012) The dynamic spatial multinomial Probit model: analysis of land use change using parcel-level data. *J Transp geogr* 24:77–88
- Yu J, de Jong R, Lee LF (2008) Quasi-maximum likelihood estimators for spatial dynamic panel data with fixed effects when both n and T are large. *J Econom* 146(1):118–134

Instrumental Variables/Method of Moments Estimation

80

Ingmar R. Prucha

Contents

80.1	Introduction	1597
80.2	A Primer on GMM Estimation	1599
80.2.1	Model Specification and Moment Conditions	1599
80.2.2	One-Step GMM Estimation	1602
80.2.3	Two-Step GMM Estimation	1604
80.3	GMM Estimation of Models with Spatial Lags	1606
80.3.1	GMM Estimation of Spatial-Autoregressive Parameter	1608
80.3.2	GMM Estimation of Regression Parameters	1609
80.3.3	Guide to Literature	1611
80.3.4	Exemplary GMM Estimators	1612
80.4	GMM Estimation of Models with Spatial Mixing	1615
80.5	Conclusion	1615
	References	1616

Abstract

The chapter discusses generalized method of moments (GMM) estimation methods for spatial models. Much of the discussion is on GMM estimation of Cliff-Ord-type models where spatial interactions are modeled in terms of spatial lags. The chapter also discusses recent developments on GMM estimation from data processes which are spatially α -mixing.

I.R. Prucha

Department of Economics, University of Maryland, College Park, MD, USA
e-mail: prucha@econ.umd.edu

80.1 Introduction

Spatial econometric models have a long history. Paelink and Klaassen (1979) may arguably be viewed as the first comprehensive volume covering spatial econometrics. Anselin (2010) provides a recent review of the development of the field of spatial econometrics over the last thirty years. Important texts include Anselin (1988), Arbia (2006), Cliff and Ord (1973, 1981), Cressie (1993), Haining (2003), and LeSage and Pace (2009).

Spatial models provide a formal expression of Tobler's (1970) first law of geography stating that "Everything is related to everything else, but near things are more related to each other." An important aspect of spatial econometrics is the focus on the explicit modeling and empirical estimation of pathways of spatial interactions. That is, an important aspect is the focus on exploring the structure of spatial interactions and not just on accounting for cross-sectional correlation in the computation of standard errors for parameter estimators.

Much of the spatial econometrics literature has focused on cross-sectional data or panel data where the time dimension is small. A reason is that in situations where the time dimension, say T , is large relative to the cross-sectional dimension, say n , we can often simply employ classical methods for the estimation of simultaneous time series models to estimate general forms of spatial interactions. If the time dimension T is one or small, estimation will only be possible if we impose some parsimonious structure on the form of spatial interactions.

The development of a formal theory of estimation of spatial models has lagged behind corresponding developments for inference from time series data. A formal theory of inference requires the use of limit theorems, such as laws of large numbers and central limit theorems. In a time series setting, there is a natural ordering of the data which can be exploited in deriving such limit theorems. In a spatial setting, there is no natural ordering of the data, which made the development of such limit theorems more challenging.

Arguably the most widely used class of spatial models consists of variants of the ones considered in Cliff and Ord (1973, 1981). In these models spatial interactions are modeled in terms of spatial lags, i.e., in terms of weighted averages of observations from neighboring units, where the weights are typically modeled as inversely related to some measure of distance. Historically, Cliff-Ord-type models have been estimated by maximum likelihood (ML) methods. (See Pace, ► Chap. 78, "Maximum Likelihood Estimation", as well as Mills and Parent, ► Chap. 79, "Bayesian MCMC Estimation".) However, one of the difficulties with ML is that the likelihood depends on the determinant of an $n \times n$ matrix, which limits its application to small and medium sample sizes due to the computational burden (unless the problem is sparse, special structure is available, etc.). Another issue was the lack of formal results concerning its asymptotic properties. In light of this, Kelejian and Prucha (1998, 1999) suggested a generalized method of moments (GMM) estimator for a spatial-autoregressive model with autoregressive

disturbances and established basic asymptotic properties for the estimator.¹ Conley (1999) considered GMM estimation within the context of α -mixing spatial processes and developed an asymptotic theory within this context.

Since those early contributions, there has been a growing literature on GMM estimation for spatially dependent data. The aim of this chapter is to provide some guidance through that literature and to provide some insights into the subtle differences in asymptotic results. Basic reasons for these differences can be found in the moment conditions employed by respective GMM estimators and whether or not an estimator is a one-step or a two-step estimator.

Owing to space limitations, the literature cited in this chapter is incomplete, and not all contributions and extensions of interest are covered. Also, the focus of this chapter is solely on GMM estimation. It does not cover maximum likelihood estimation or testing procedures (apart from Wald tests that can be constructed in the usual way based on results for the asymptotic distribution of GMM estimators). Also, the chapter does not cover inference for processes where cross-sectional dependence is implied by common factors.

Finally, while spatial models have a long history in geography and regional science, space is not limited to geographic space. Spatial models may more generally be viewed as a class of cross-sectional interaction models, with applications ranging from growth convergence among regions to social interactions between agents.

Section 80.2 of the chapter contains a brief and intuitive primer on GMM estimation to provide some background. Readers familiar with GMM estimation may wish to skip this section. Section 80.3 considers GMM estimation of models with spatial lags, and Sect. 80.4 considers GMM estimation for a general class of spatially mixing processes.

80.2 A Primer on GMM Estimation

80.2.1 Model Specification and Moment Conditions

Suppose the data are generated from a model

$$f(y_{in}, z_{in}, \theta_0) = u_{in}, \quad i = 1, \dots, n \quad (80.1)$$

where y_{in} denotes the dependent variable corresponding to unit i , z_{in} is a vector of explanatory variables, u_{in} is a disturbance term, θ_0 is the $K \times 1$ unknown parameter vector, and $f(\cdot)$ is a known function. The above formulation is fairly general and contains typical Cliff and Ord (1973) spatial models – possibly after some

¹Lee (2004) gives, to the best of our knowledge, first formal results for the maximum likelihood estimator of a spatial-autoregressive model. The maintained assumptions are similar to those introduced in Kelejian and Prucha (1998, 1999).

transformation to remove correlation in the disturbance term – as a special case. Additionally assume the availability of a $1 \times P$ vector of instruments h_{in} and let w_{in} be the vector of all observable variables, including instruments, pertaining to the i th unit. For simplicity of presentation, we assume in the following that the disturbances are i.i.d. $(0, \sigma^2)$ and that the instruments are non-stochastic while noting that both assumptions can be relaxed.

We also note that in allowing for the variables to depend on the sample size, we accommodate spatial lags. As an example, the explanatory variables could be of the form $z_{in} = [x_i, \bar{x}_{in}, \bar{y}_{in}]$ where x_i is some exogenous explanatory variable, and $\bar{x}_{in} = \sum_j m_{ij}x_j$ and $\bar{y}_{in} = \sum_j m_{ij}y_j$ are spatial lags (where the m_{ij} denote spatial weights with $m_{ii} = 0$). However, to simplify notation for this primer, we will suppress the index n in the following.

Now suppose that we have a $S \times 1$ vector of sample moments

$$\mathbf{q}_n(\theta) = \mathbf{q}_n(w_1, \dots, w_n, \theta) = \begin{bmatrix} q_{1,n}(w_1, \dots, w_n, \theta) \\ \vdots \\ q_{S,n}(w_1, \dots, w_n, \theta) \end{bmatrix} \quad (80.2)$$

with $S \geq K$, and suppose that

$$E\mathbf{q}_n(w_1, \dots, w_n, \theta) = 0 \quad \text{if and only if} \quad \theta = \theta_0 \quad (80.3)$$

The basic idea underlying the GMM methodology is to estimate θ_0 by, say, $\tilde{\theta}_n$ such that $\mathbf{q}_n(w_1, \dots, w_n, \tilde{\theta}_n)$ is “close to zero” in the sense that a quadratic form of the sample moment vector is close to zero. More specifically, let \mathbf{Y}_n be some $S \times S$ symmetric positive semidefinite weighting matrix, then the corresponding GMM estimator is defined as

$$\tilde{\theta}_n = \arg \min_{\theta} \{ \mathbf{q}_n(w_1, \dots, w_n, \theta)' \mathbf{Y}_n \mathbf{q}_n(w_1, \dots, w_n, \theta) \} \quad (80.4)$$

A special case arises if the number of moments equals the number of unknown parameters, i.e., if $S = K$. In this case $\tilde{\theta}_n$ can typically be found as a solution to the moment condition, i.e., $\mathbf{q}_n(w_1, \dots, w_n, \tilde{\theta}_n) = 0$. Of course, in this case the weighting matrix \mathbf{Y}_n becomes irrelevant.

The classical GMM literature exploits “linear” moment conditions of the form

$$E \left\{ n^{-1} \sum_{i=1}^n h'_{ip} u_i \right\} = 0 \quad (80.5)$$

which clearly holds since $Eh'_{ip} u_i = h'_{ip} Eu_i = 0$ under the maintained assumptions. The spatial literature frequently also considers “quadratic” moment conditions. Let $A_q = (a_{ijq})$ be some $n \times n$ matrix with $tr(A_q) = 0$, and assume for ease of

exposition that A_q is non-stochastic. Then the quadratic moment conditions considered in the spatial literature are of the form

$$E \left\{ n^{-1} \sum_{i=1}^n \sum_{j=1}^n a_{ijq} u_i u_j \right\} = 0 \quad (80.6)$$

which clearly holds under maintained assumptions.²

Depending on the functional form of $f(\cdot)$, the number of moment conditions, the number of parameters, etc., the computation of the GMM estimator $\tilde{\theta}_n$ defined by Eq. (80.4) may be numerically challenging. Now let $\theta_0 = [\rho'_0, \delta'_0]'$ and suppose the sample moment vector in Eq. (80.2) can be decomposed into

$$\mathbf{q}_n(w_1, \dots, w_n, \theta) = \begin{bmatrix} \mathbf{q}_n^\rho(w_1, \dots, w_n, \rho, \delta) \\ \mathbf{q}_n^\delta(w_1, \dots, w_n, \rho, \delta) \end{bmatrix} \quad (80.7)$$

such that

$$E \mathbf{q}_n^\rho(w_1, \dots, w_n, \rho, \delta_0) = 0 \quad \text{if and only if} \quad \rho = \rho_0 \quad (80.8)$$

$$E \mathbf{q}_n^\delta(w_1, \dots, w_n, \rho_0, \delta) = 0 \quad \text{if and only if} \quad \delta = \delta_0 \quad (80.9)$$

and that some easily computable initial estimator, say $\check{\delta}_n$, for δ_0 is available. In this case we may consider the following GMM estimator for ρ_0 corresponding to some weighting matrix $\mathbf{Y}_n^{\rho\rho}$:

$$\hat{\rho}_n = \arg \min_{\rho} \left\{ \mathbf{q}_n^\rho(w_1, \dots, w_n, \rho, \check{\delta}_n)' \mathbf{Y}_n^{\rho\rho} \mathbf{q}_n^\rho(w_1, \dots, w_n, \rho, \check{\delta}_n) \right\} \quad (80.10)$$

Of course, utilizing $\hat{\rho}_n$ we may further consider the following GMM estimator for δ_0 corresponding to some weight matrix $\mathbf{Y}_n^{\delta\delta}$:

$$\hat{\delta}_n = \arg \min_{\delta} \left\{ \mathbf{q}_n^\delta(w_1, \dots, w_n, \hat{\rho}_n, \delta)' \mathbf{Y}_n^{\delta\delta} \mathbf{q}_n^\delta(w_1, \dots, w_n, \hat{\rho}_n, \delta) \right\} \quad (80.11)$$

GMM estimators like $\tilde{\theta}_n$ in Eq. (80.4) are often referred to as one-step estimators. Estimators like $\hat{\rho}_n$ and $\hat{\delta}_n$ in Eqs. (80.10) and (80.11) above, where the sample moments depend on some initial estimator, are often referred to as two-step estimators.

Given the moment conditions are valid, we would expect the most efficient one-step estimator to be more efficient than the most efficient two-step estimators. However, as

²Let $u = [u_1, \dots, u_n]'$, then the above moment condition can be rewritten as $E[n^{-1}u'A_qu] = \text{tr}[n^{-1}A_qEu'u'] = n^{-1}\sigma^2\text{tr}(A_q) = 0$, since under the maintained assumptions $Eu'u' = \sigma^2 I_n$.

usual, there are trade-offs. One trade-off is in terms of computation. As remarked previously, for small sample sizes ML is available as an alternative to GMM. For large sample sizes, statistical efficiency may be less important than computational efficiency and feasibility, and thus the use of two-step GMM estimators may be attractive. Also, Monte Carlo studies suggest that in many situations, the loss of efficiency may be relatively small. Another trade-off is that the misspecification of one moment condition will typically result in inconsistent estimates of all model parameters.

In the following we provide some basic results for the limiting distribution of one-step and two-step GMM estimators as background for our discussion of specific GMM estimators for respective spatial models.

80.2.2 One-Step GMM Estimation

The usual approach to deriving the limiting distribution of GMM estimators is to manipulate the score of the objective function by expanding the sample moment vector around the true parameter, using a Taylor expansion. Applying this approach to Eq. (80.4), and assuming that typical regularity conditions hold, yields

$$n^{1/2}(\hat{\theta}_n - \theta_0) = -[\mathbf{G}'\mathbf{Y}\mathbf{G}]^{-1}\mathbf{G}'\mathbf{Y}\left[n^{1/2}\mathbf{q}_n(\theta_0)\right] + o_p(1) \quad (80.12)$$

with $\mathbf{G} = p \lim_{n \rightarrow \infty} \partial \mathbf{q}_n(\theta_0) / \partial \theta$ and $\mathbf{Y} = p \lim_{n \rightarrow \infty} \mathbf{Y}_n$. Now suppose for a moment that it can be shown that

$$n^{1/2}\mathbf{q}_n(\theta_0) \xrightarrow{d} N(0, \Psi) \quad (80.13)$$

where Ψ is some positive definite matrix. Then

$$n^{1/2}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N[0, \Phi] \quad (80.14)$$

with

$$\Phi = [\mathbf{G}'\mathbf{Y}\mathbf{G}]^{-1}\mathbf{G}'\mathbf{Y}\Psi\mathbf{Y}\mathbf{G}[\mathbf{G}'\mathbf{Y}\mathbf{G}]^{-1}$$

From this it is seen that if we choose $\mathbf{Y}_n = \hat{\Psi}_n^{-1}$ where $\Psi = p \lim_{n \rightarrow \infty} \hat{\Psi}_n$, the variance-covariance simplifies to

$$\Phi = [\mathbf{G}'\Psi^{-1}\mathbf{G}]^{-1}$$

Since $[\mathbf{G}'\mathbf{Y}\mathbf{G}]^{-1}\mathbf{G}'\mathbf{Y}\Psi\mathbf{Y}\mathbf{G}[\mathbf{G}'\mathbf{Y}\mathbf{G}]^{-1} - [\mathbf{G}'\Psi^{-1}\mathbf{G}]^{-1}$ is positive semidefinite, it follows that using for the weighting matrix \mathbf{Y}_n , a consistent estimator of the inverse of the limiting variance-covariance matrix Ψ of the sample moment vector yields the efficient GMM estimator.

As remarked above, for spatial estimators the sample moment vector will typically be composed of linear and quadratic moment conditions of the form given in Eqs. (80.4) and (80.5). Thus, in order to establish Eq. (80.13), we need a central limit theorem (CLT) for linear quadratic forms. Kelejian and Prucha (2001) introduced such a theorem for a single linear quadratic form under assumptions useful for spatial models. The generalization to vectors of linear quadratic forms is given in Kelejian and Prucha (2010). To provide some insight into the expressions for the asymptotic variance-covariance matrix Ψ associated with the sample moment vector underlying the spatial GMM estimators below, we next give a version of that CLT.

Theorem 1

For $r = 1, \dots, m$ let $A_{r,n} = (a_{ijr})_{i,j=1,\dots,n}$ be a $n \times n$ non-stochastic symmetric real matrix with $\sup_{1 \leq j \leq n, n \geq 1} \sum_{i=1}^n |a_{ijr}| < \infty$, and let $a_r = (a_{1r}, \dots, a_{nr})'$ be a $n \times 1$ non-stochastic real vector with $\sup_n n^{-1} \sum_{i=1}^n |a_{ir}|^{\delta_1} < \infty$ for some $\delta_1 > 2$. Let $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n)'$ be a $n \times 1$ random vector with the ε_i distributed totally independent with $E\varepsilon_i = 0$, $E\varepsilon_i^2 = \sigma_i^2$, and $\sup_{1 \leq i \leq n, n \geq 1} E|\varepsilon_i|^{\delta_2} < \infty$ for some $\delta_2 > 4$. Consider the $m \times 1$ vector of linear quadratic forms $V_n = [Q_{1n}, \dots, Q_{mn}]'$ with

$$Q_{rn} = \varepsilon' A_r \varepsilon + a'_r \varepsilon = \sum_{i=1}^n \sum_{j=1}^n a_{ijr} \varepsilon_i \varepsilon_j + \sum_{i=1}^n a_{ir} \varepsilon_i$$

Let $\mu_{V_n} = EV_n = [\mu_{Q_1}, \dots, \mu_{Q_m}]'$ and $\Sigma_{V_n} = [\sigma_{Q_{rs}}]_{r,s=1,\dots,m}$ denote the mean and VC matrix of V_n , respectively, then

$$\begin{aligned} \mu_{Q_r} &= \sum_{i=1}^n a_{iir} \sigma_i^2 \\ \sigma_{Q_{rs}} &= 2 \sum_{i=1}^n \sum_{j=1}^n a_{ijr} a_{ijs} \sigma_i^2 \sigma_j^2 + \sum_{i=1}^n a_{ir} a_{is} \sigma_i^2 \\ &\quad + \sum_{i=1}^n a_{iir} a_{iis} \left[\mu_i^{(4)} - 3\sigma_i^4 \right] + \sum_{i=1}^n (a_{ir} a_{iis} + a_{is} a_{iir}) \mu_i^{(3)} \end{aligned}$$

with $\mu_i^{(3)} = E\varepsilon_i^3$ and $\mu_i^{(4)} = E\varepsilon_i^4$. Furthermore, given that $n^{-1} \lambda_{\min}(\Sigma_{V_n}) \geq c$ for some $c > 0$, then

$$\Sigma_{V_n}^{-1/2} (V_n - \mu_{V_n}) \xrightarrow{d} N(0, I_m)$$

and thus,

$$n^{-1/2} (V_n - \mu_{V_n}) \sim AN(0, n^{-1} \Sigma_{V_n})$$

Remark

Note that the mean μ_{Q_r} of Q_{rn} is zero if $a_{ir} = 0$; if the ε_i are homoskedastic, i.e., $\sigma_i^2 = \sigma^2$, then $tr(A_r) = \sum_{i=1}^n a_{ir} = 0$ suffices for the mean to be zero. Next, note that the first two terms in the expression for the covariance $\sigma_{Q_{rs}}$ between Q_{rn} and Q_{sn} can be written more compactly as $2tr(A_r \Sigma A_s \Sigma) + a_r' \Sigma a_s$ with $\Sigma = diag(\sigma_i^2)$. Also note that if $a_{ir} = a_{is} = 0$, then the last two terms drop out from the expression for covariance. Observe further that under normality, the last two terms are always equal to zero.

80.2.3 Two-Step GMM Estimation

The derivation of the limiting distribution two-step of GMM estimators is a bit more delicate. The usual approach to deriving the limiting distribution of two-step GMM estimators is to manipulate the score of the objective function by expanding the sample moment vector around the true parameter, using a Taylor expansion. Consider in particular the two-step GMM estimators for ρ_0 defined in Eq. (80.10). Applying this approach, and assuming typical regularity conditions, yields

$$\begin{aligned} & n^{1/2}(\hat{\rho}_n - \rho_0) \\ &= -[(\mathbf{G}^{\rho\rho})' \mathbf{Y}^{\rho\rho} \mathbf{G}^{\rho\rho}]^{-1} (\mathbf{G}^{\rho\rho})' \mathbf{Y}^{\rho\rho} \left[n^{1/2} \mathbf{q}_n^\rho(\rho_0, \delta_0) + \mathbf{G}^{\rho\delta} n^{1/2} (\check{\delta}_n - \delta_0) \right] + o_p(1) \end{aligned} \quad (80.15)$$

where $\mathbf{G}^{\rho\rho} = p \lim_{n \rightarrow \infty} \partial \mathbf{q}_n^\rho(\rho_0, \delta_0) / \partial \rho$, $\mathbf{G}^{\rho\delta} = p \lim_{n \rightarrow \infty} \partial \mathbf{q}_n^\rho(\rho_0, \delta_0) / \partial \delta$, and $\mathbf{Y}^{\rho\rho} = p \lim_{n \rightarrow \infty} \mathbf{Y}_n^{\rho\rho}$. From Eq. (80.1) we see that in general the limiting distribution of $\hat{\rho}_n$ will depend on the limiting distribution of $\check{\delta}_n$, unless $\mathbf{G}^{\rho\delta} = 0$, in which case we refer to $\check{\delta}_n$ as a nuisance parameter. It turns out that if ρ_0 denotes the spatial-autoregressive parameters in the disturbance process and δ_0 the vector of regression parameters, then for typical estimators $\mathbf{G}^{\rho\delta} \neq 0$. In many cases the estimator $\check{\delta}_n$ will be asymptotically linear in the sense that

$$n^{1/2}(\check{\delta}_n - \delta_0) = n^{-1/2} \mathbf{T}'_n \mathbf{u}_n + o_p(1) \quad (80.16)$$

where \mathbf{T}_n is a non-stochastic $n \times k_\delta$ matrix, where k_δ is the dimension of δ_0 , and where $\mathbf{u}_n = (u_1, \dots, u_n)'$. Now define

$$\mathbf{q}_{*n}^\rho(\rho_0, \delta_0) = \mathbf{q}_n^\rho(\rho_0, \delta_0) + n^{-1} \mathbf{G}^{\rho\delta} \mathbf{T}'_n \mathbf{u}_n \quad (80.17)$$

then Eq. (80.15) can be rewritten as

$$n^{1/2}(\hat{\rho}_n - \rho_0) = -[(\mathbf{G}^{\rho\rho})' \mathbf{Y}^{\rho\rho} \mathbf{G}^{\rho\rho}]^{-1} (\mathbf{G}^{\rho\rho})' \mathbf{Y}^{\rho\rho} \left[n^{1/2} \mathbf{q}_{*n}^\rho(\rho_0, \delta_0) \right] + o_p(1) \quad (80.18)$$

Now suppose that

$$n^{1/2} \mathbf{q}_{*n}^\rho(\rho_0, \delta_0) \xrightarrow{d} N(0, \Psi_*^{\rho\rho}) \quad (80.19)$$

where $\Psi_*^{\rho\rho}$ is some positive definite matrix. Then

$$n^{1/2}(\hat{\rho}_n - \rho_0) d \xrightarrow{d} N[0, \Phi_*^{\rho\rho}] \quad (80.20)$$

with

$$\Phi_*^{\rho\rho} = [(\mathbf{G}^{\rho\rho})' \mathbf{Y}^{\rho\rho} \mathbf{G}^{\rho\rho}]^{-1} (\mathbf{G}^{\rho\rho})' \mathbf{Y}^{\rho\rho} \Psi_*^{\rho\rho} \mathbf{Y}^{\rho\rho} \mathbf{G}^{\rho\rho} [(\mathbf{G}^{\rho\rho})' \mathbf{Y}^{\rho\rho} \mathbf{G}^{\rho\rho}]^{-1}$$

From this it is seen that if we choose $\mathbf{Y}_n^{\rho\rho} = (\tilde{\Psi}_{*n}^{\rho\rho})^{-1}$ where $\Psi_*^{\rho\rho} = p \lim_{n \rightarrow \infty} \tilde{\Psi}_{*n}^{\rho\rho}$, then variance-covariance simplifies to

$$\Phi_*^{\rho\rho} = [(\mathbf{G}^{\rho\rho})' (\Psi_*^{\rho\rho})^{-1} \mathbf{G}^{\rho\rho}]^{-1}$$

Therefore, using for the weighting matrix $\mathbf{Y}_n^{\rho\rho}$, a consistent estimator for the inverse of the limiting variance-covariance matrix $\Psi_*^{\rho\rho}$ yields the efficient two-step GMM estimator.

Suppose Eq. (80.13) holds and

$$\Psi = \begin{bmatrix} \Psi^{\rho\rho} & \Psi^{\rho\delta} \\ \Psi^{\delta\rho} & \Psi^{\delta\delta} \end{bmatrix}$$

then the limiting distribution of the sample moment vector $\mathbf{q}_n^\rho(\rho_0, \delta_0)$ evaluated at the true parameter values is given by

$$n^{1/2} \mathbf{q}_n^\rho(\rho_0, \delta_0) d \rightarrow N(0, \Psi^{\rho\rho}) \quad (80.21)$$

It is important to note that in light of Eq. (80.17) in general $\Psi_*^{\rho\rho} \neq \Psi^{\rho\rho}$, unless $\mathbf{G}^{\rho\delta} = 0$, and that in general $\Psi_*^{\rho\rho}$ will depend on \mathbf{T}_n , which in turn will depend on the employed estimator δ_n . In other words, unless $\mathbf{G}^{\rho\delta} = 0$, for a two-step GMM estimator, we cannot simply use the variance-covariance matrix $\Psi^{\rho\rho}$ of the sample moment vector $\mathbf{q}_n^\rho(\rho_0, \delta_0)$, rather we need to work with the variance-covariance matrix $\Psi_*^{\rho\rho}$.

We next illustrate the difference between $\Psi^{\rho\rho} = (\psi_{rs}^{\rho\rho})$ and $\Psi_*^{\rho\rho} = (\psi_{*rs}^{\rho\rho})$ for the important special case where the moment conditions are quadratic and u_i is i.i.d. $N(0, \sigma^2)$. For simplicity assume that

$$\mathbf{q}_n^\rho(\rho_0, \delta_0) = n^{-1} \begin{bmatrix} \sum_{i=1}^n \sum_{j=1}^n a_{ij1} u_i u_j \\ \sum_{i=1}^n \sum_{j=1}^n a_{ij2} u_i u_j \end{bmatrix}$$

Now, for $r = 1, 2$, let a_{ir} denote the (i, r) -th element of $\mathbf{G}^{\rho\delta} \mathbf{T}'_n$, then by Eq. (80.17)

$$\mathbf{q}_{*n}^\rho(\rho_0, \delta_0) = n^{-1} \begin{bmatrix} \sum_{i=1}^n \sum_{j=1}^n a_{ij1} u_i u_j + \sum_{i=1}^n a_{i1} u_i \\ \sum_{i=1}^n \sum_{j=1}^n a_{ij2} u_i u_j + \sum_{i=1}^n a_{i2} u_i \end{bmatrix}$$

It then follows from Theorem 1 that

$$\psi_{rs}^{\rho\rho} = 2\sigma^4 \sum_{i=1}^n \sum_{j=1}^n a_{ijr} a_{ijs}$$

but

$$\psi_{*rs}^{\rho\rho} = 2\sigma^4 \sum_{i=1}^n \sum_{j=1}^n a_{ijr} a_{ijs} + \sigma^2 \sum_{i=1}^n a_{ir} a_{is}$$

We emphasize that the a_{ir} and a_{is} in the last sum on the r.h.s. for the expression for $\psi_{*rs}^{\rho\rho}$ depend on what estimator $\check{\delta}_n$ is employed in the sample moment vector $\mathbf{q}_n^\rho(\rho, \check{\delta}_n)$ used to form the objective function for the two-step GMM estimator $\hat{\rho}_n$ defined in Eq. (80.10). It is for this reason that in the literature on two-step GMM estimation, users are often advised to follow a specific sequence of steps, to ensure the proper estimation of respective variance-covariance matrices.

80.3 GMM Estimation of Models with Spatial Lags

As remarked in the introduction, arguably the most widely used class of spatial models are variants of the ones considered in Cliff and Ord (1973, 1981), which build on the fundamental contribution of Whittle (1954). In these models, spatial interactions are modeled in terms of spatial lags. In particular, consider the following Cliff-Ord-type model relating a cross section of n spatial units:

$$\begin{aligned} \mathbf{y}_n &= \mathbf{X}_n \beta_{0n} + \lambda_{0n} \mathbf{W}_n \mathbf{y}_n + \mathbf{u}_n \\ &= \mathbf{Z}_n \delta_{0n} + \mathbf{u}_n \\ \mathbf{u}_n &= \rho_{0n} \mathbf{M}_n \mathbf{u}_n + \varepsilon_n \end{aligned} \tag{80.22}$$

where $\mathbf{Z}_n = [\mathbf{X}_n, \mathbf{W}_n \mathbf{y}_n]$ and $\delta_{0n} = [\beta'_{0n}, \lambda_{0n}]'$. Here $\mathbf{y}_n = (y_{1,n}, \dots, y_{n,n})'$ is the $n \times 1$ vector of the dependent variable, $\mathbf{X}_n = (x_{ik,n})$ is the $n \times K$ matrix of the non-stochastic exogenous regressors, $\mathbf{W}_n = (w_{ij,n})$ and $\mathbf{M}_n = (m_{ij,n})$ are $n \times n$ observed non-stochastic weight matrices with zero diagonal elements, $\mathbf{u}_n = (u_{1,n}, \dots, u_{n,n})'$ is the $n \times 1$ vector of regression disturbances, and $\varepsilon_n = (\varepsilon_{1,n}, \dots, \varepsilon_{n,n})'$ is an $n \times 1$ vector of innovations. The vectors $\bar{\mathbf{y}}_n = (\bar{y}_{1,n}, \dots, \bar{y}_{n,n})' = \mathbf{W}_n \mathbf{y}_n$ and $\bar{\mathbf{u}}_n = (\bar{u}_{1,n}, \dots, \bar{u}_{n,n})' = \mathbf{M}_n \mathbf{u}_n$ represent spatial lags, the scalars λ_{0n} and ρ_{0n} denote the corresponding true parameters, typically referred to as spatial-autoregressive parameters, and β_{0n} is a $k \times 1$ true parameter vector. In analogy to the time series literature, the above model is often referred to as a spatial-autoregressive autoregressive (1,1) model, for short an SARAR(1,1) model.

In the above formulation, all data vectors and matrices, as well as all parameters are allowed to depend on the sample size n , i.e., to form triangular arrays. To see why this is necessary, consider, e.g., the i th elements of the spatial lag $\bar{\mathbf{y}}_n = \mathbf{W}_n \mathbf{y}_n$, which is given by

$$\bar{y}_{i,n} = \sum_{j=1}^n w_{ij,n} y_{j,n}$$

From this it is obvious that even if the weights $w_{ij,n}$ do not depend on n , the weighted average $\bar{y}_{i,n}$ and thus $y_{i,n}$ will depend on n . In allowing for the elements of \mathbf{X}_n to depend on n , we allow implicitly for some of the regressors to be spatial lags, e.g., the regressor matrix could be of the form $\mathbf{X}_n = [\mathbf{x}_{1,n}, \mathbf{W}_n \mathbf{x}_{1,n}, \dots]$. In allowing for the elements of the spatial weight matrices to depend on n , we allow implicitly for normalized spatial weight matrices, as is frequently the case in applications. In allowing also for the parameters to depend on n allows us to assume a common parameter space for all sample sizes; see Kelejian and Prucha (2010) for a more detailed discussion. For simplicity of notation we will, for the most part, drop again subscripts n in the following.

The spatial model (80.22) represents a system of n simultaneous equations. The reduced form of the model is given by

$$\mathbf{y} = (\mathbf{I} - \lambda_0 \mathbf{W})^{-1} \mathbf{X} \beta_0 + (\mathbf{I} - \lambda_0 \mathbf{W})^{-1} (\mathbf{I} - \rho_0 \mathbf{M})^{-1} \varepsilon \quad (80.23)$$

If $\varepsilon \sim N(0, \sigma^2 \mathbf{I})$, then clearly $\mathbf{y} \sim N(\mu_y, \Omega_y)$ with

$$\begin{aligned} \mu_y &= (\mathbf{I} - \lambda_0 \mathbf{W})^{-1} \mathbf{X} \beta_0, \\ \Omega_y &= \sigma^2 (\mathbf{I} - \lambda_0 \mathbf{W})^{-1} (\mathbf{I} - \rho_0 \mathbf{M})^{-1} (\mathbf{I} - \rho_0 \mathbf{M}')^{-1} (\mathbf{I} - \lambda_0 \mathbf{W}')^{-1} \end{aligned}$$

From this we see that while it is easy to write down the log-likelihood function for model (80.22), the computation of the ML estimator is challenging or non-feasible for larger sample sizes n . The reason is that it requires the computation of

the determinant of the $n \times n$ matrices $\mathbf{I} - \lambda_0 \mathbf{W}$ and $\mathbf{I} - \rho_0 \mathbf{M}$, which is taxing for large n unless the spatial weight matrices have structure that can be exploited.

Our discussions will also utilize the following spatial Cochrane-Orcutt transformation of Eq. (80.22):

$$\mathbf{y}_*(\rho_0) = \mathbf{Z}_*(\rho_0)\delta_0 + \varepsilon \quad (80.24)$$

where $\mathbf{y}_*(\rho_0) = \mathbf{y} - \rho_0 \mathbf{M}\mathbf{y}$ and $\mathbf{Z}_*(\rho_0) = \mathbf{Z} - \rho_0 \mathbf{M}\mathbf{Z}$. The transformed model is readily obtained by pre-multiplying Eq. (80.22) by $\mathbf{I} - \rho_0 \mathbf{M}$.

80.3.1 GMM Estimation of Spatial-Autoregressive Parameter

Motived by the potential numerical problems in computing the ML estimator for larger sample sizes, Kelejian and Prucha (1998, 1999) introduced an alternative GMM estimation approach which remains feasible even for large sample sizes and full spatial weight matrices.³ (See Pace, ▶ Chap. 78, “Maximum Likelihood Estimation”.) Another motivation was that at the time there were no formal results available regarding the consistency and asymptotic normality of the ML estimator for the above model.

The GMM estimation approach put forward in Kelejian and Prucha (1998, 1999) employs the following simple quadratic moment conditions, based on the assumption that the ε_i are i.i.d. $(0, \sigma^2)$:

$$En^{-1}\varepsilon'\varepsilon = \sigma^2, \quad En^{-1}\bar{\varepsilon}'\bar{\varepsilon} = \sigma^2n^{-1}tr(\mathbf{M}'\mathbf{M}), \quad En^{-1}\bar{\varepsilon}'\varepsilon = 0$$

with $\bar{\varepsilon} = \mathbf{M}\varepsilon$. Substituting out σ^2 yields the following two quadratic moment conditions:

$$En^{-1}\varepsilon'\mathbf{A}_1\varepsilon = 0, \quad En^{-1}\varepsilon'\mathbf{A}_2\varepsilon = 0 \quad (80.25)$$

with⁴

$$\mathbf{A}_1 = \mathbf{M}'\mathbf{M} - n^{-1}tr(\mathbf{M}'\mathbf{M})\mathbf{I}, \quad \mathbf{A}_2 = \mathbf{M} \quad (80.26)$$

We note that for the weight matrices in Eq. (80.5), we have $tr(\mathbf{A}_q) = 0$ for $q = 1, 2$, but $diag(\mathbf{A}_1) \neq 0$. Kelejian and Prucha (2010) relax the assumption that the innovations are homoskedastic and allow for heteroskedasticity of unknown form.

³Recall, e.g., that there are more than 33,000 zip codes in the U.S.

⁴To obtain the estimator of Kelejian and Prucha (1998, 1999) the matrix \mathbf{A}_1 has to be scaled by $v = 1 / [1 + [n^{-1}tr(\mathbf{M}'\mathbf{M})]^2]$. Of course, the scaling factor only comes into play if the moment conditions are not optimally weighted, as was the case in the early literature.

More specifically, they consider the case where the ε_i are independently distributed $(0, \sigma_i^2)$ with σ_i^2 unknown.⁵ For this case they consider the following modified version of the above moment conditions where

$$\mathbf{A}_1 = \mathbf{M}'\mathbf{M} - n^{-1}\text{diag}(\mathbf{M}'\mathbf{M}), \quad \mathbf{A}_2 = \mathbf{M} \quad (80.27)$$

Note that in this specification, $\text{diag}(\mathbf{A}_q) = 0$ for $q = 1, 2$. Given this, the moment conditions in Eq. (80.25) continue to hold since $E\varepsilon'\mathbf{A}_q\varepsilon = \sum_{i=1}^n a_{q,ii}\sigma_i^2 = 0$. From this we see that, in general, moment conditions that employ weight matrices with $a_{q,ii} = 0$ and not just $\text{tr}(\mathbf{A}_q) = \sum_{i=1}^n a_{q,ii} = 0$ are robust against heteroskedasticity.

Of course, the above setup can be generalized to the case where we have S_ρ quadratic moment conditions ($q = 1, \dots, S_\rho$):

$$En^{-1}\varepsilon'\mathbf{A}_q\varepsilon = 0 \quad (80.28)$$

In light of Eq. (80.22) those moment conditions can be written equivalently as ($q = 1, \dots, S_\rho$):

$$En^{-1}\mathbf{u}'(\mathbf{I} - \rho_0\mathbf{M}')\mathbf{A}_q(\mathbf{I} - \rho_0\mathbf{M})\mathbf{u} = 0 \quad (80.29)$$

Now let $\check{\delta}$ be some initial estimator for δ_0 and let $\check{\mathbf{u}} = \mathbf{y} - \mathbf{Z}\check{\delta}$. Then we can formulate the following corresponding sample moment vector:

$$\mathbf{q}_n^\rho(\rho, \check{\delta}) = n^{-1} \begin{bmatrix} \check{\mathbf{u}}'(\mathbf{I} - \rho\mathbf{M}')\mathbf{A}_1(\mathbf{I} - \rho\mathbf{M})\check{\mathbf{u}} \\ \vdots \\ \check{\mathbf{u}}'(\mathbf{I} - \rho\mathbf{M}')\mathbf{A}_{S_\rho}(\mathbf{I} - \rho\mathbf{M})\check{\mathbf{u}} \end{bmatrix} \quad (80.30)$$

Furthermore, as in Eq. (80.10), the class of corresponding two-step GMM estimators is then given by

$$\hat{\rho} = \arg \min_{\rho} \{ \mathbf{q}_n^\rho(\rho, \check{\delta})' \mathbf{Y}_n^{\rho\rho} \mathbf{q}_n^\rho(\rho, \check{\delta}) \} \quad (80.31)$$

where $\mathbf{Y}_n^{\rho\rho}$ is a weighting matrix. As discussed in Sect. 80.2.3, the efficient choice for $\mathbf{Y}_n^{\rho\rho}$ will generally depend on the estimator $\check{\delta}$ employed in the estimation of the disturbances.

⁵Lin and Lee (2010) also allow for heteroskedastic innovations for model (80.22) with $\rho_0 = 0$.

80.3.2 GMM Estimation of Regression Parameters

In order to motivate the GMM estimator for the regression parameters δ_0 , we note that the best instruments for the r.h.s. variables of model (80.22) and (80.24) are the conditional means. Since \mathbf{X} and \mathbf{MX} are non-stochastic (and their own best instruments), we can focus on the spatial lags \mathbf{Wy} and \mathbf{MWy} . The best instruments are given \mathbf{WEy} and \mathbf{MWEy} with

$$\mathbf{Ey} = (\mathbf{I} - \lambda_0 \mathbf{W})^{-1} \mathbf{X} \beta_0 = \sum_{l=1}^{\infty} \lambda_0^l \mathbf{W}^l \mathbf{X} \beta_0 \quad (80.32)$$

given that spectral radius of $\lambda_0 \mathbf{W}$ is less than one. To avoid issues associated with the computation of the inverse of the $n \times n$ matrix of $\mathbf{I} - \lambda_0 \mathbf{W}$, Kelejian and Prucha (1998, 1999) suggest the use of an approximation of the best instruments. More specifically, in light of the last expression in Eq. (80.32), they suggest using a set of instruments \mathbf{H} which contains, say, $\mathbf{X}, \mathbf{MX}, \mathbf{MWX}, \dots, \mathbf{MW}^p \mathbf{X}$, and to compute approximators of the best instruments from a regression of the r.h.s. variables against \mathbf{H} .

For the untransformed model, this is equivalent to considering the moment condition $En^{-1} \mathbf{H}' u = 0$. Of course, the corresponding GMM estimator is just the two-stage least squares (2SLS) estimator. For the transformed model (80.24), the moment condition would be

$$En^{-1/2} \mathbf{H}' e = 0 \quad (80.33)$$

Now let $\check{\rho}$ be some estimator for ρ_0 , then we can formulate the following corresponding sample moment vector:

$$\mathbf{q}^\delta(\check{\rho}, \delta) = n^{-1/2} \mathbf{H}' [\mathbf{y}_*(\check{\rho}) - \mathbf{Z}_*(\check{\rho}) \delta] \quad (80.34)$$

Under homoskedasticity the variance-covariance matrix of the moment vector $\mathbf{q}^\delta(\rho_0, \delta_0) = n^{-1/2} \mathbf{H}' e$ is given by $\sigma^2 n^{-1} \mathbf{H}' \mathbf{H}$, which motivates the following two-step GMM estimators for δ_0 :

$$\widehat{\delta} = \arg \min_{\delta} \left\{ \mathbf{q}_n^\delta(\check{\rho}, \delta)' \mathbf{Y}_n^{\delta\delta} \mathbf{q}_n^\delta(\check{\rho}, \delta) \right\} \quad (80.35)$$

with $\mathbf{Y}_n^{\delta\delta} = [n^{-1} \mathbf{H}' \mathbf{H}]^{-1}$. By observing that the quadratic form on the r.h.s. of Eq. (80.3) is just $[\mathbf{y}_*(\check{\rho}) - \mathbf{Z}_*(\check{\rho}) \delta]' \mathbf{H} (\mathbf{H}' \mathbf{H})^{-1} \mathbf{H}' [\mathbf{y}_*(\check{\rho}) - \mathbf{Z}_*(\check{\rho}) \delta]$, apart from some scaling factors, we see that the estimator defined by Eq. (80.35) is just the 2SLS estimator applied to the transformed model (80.24) with ρ_0 replaced by $\check{\rho}$, i.e.,

$$\widehat{\delta} = [\widehat{\mathbf{Z}}_*(\check{\rho})' \mathbf{Z}_*(\check{\rho})]^{-1} \widehat{\mathbf{Z}}_*(\check{\rho})' \mathbf{y}_*(\check{\rho}) \quad (80.36)$$

where $\widehat{\mathbf{Z}}_*(\check{\rho}) = \mathbf{H}(\mathbf{H}'\mathbf{H})^{-1}\mathbf{H}'\mathbf{Z}_*(\check{\rho})$. This estimator has been called the feasible generalized spatial two-stage least squares (FGS2SLS) estimator.

80.3.3 Guide to Literature

The above sections discussed basic ideas concerning moment conditions that can be exploited by GMM estimators for spatial Cliff-Ord-type models. Since the late 1990s, a considerable body of literature has developed regarding the GMM estimation of Cliff-Ord-type models. In the following we provide references to some of that literature. Naturally, given space limitations, the list of references is incomplete. Also, the list will focus on theoretical contributions and will not cover corresponding empirical work.⁶

By employing an approximation of the best instruments, the FGS2SLS estimator of Kelejian and Prucha (1998, 1999) has the advantage of remaining computational feasible even for very large sample sizes since its formulation does not involve the computation of the inverse of the $n \times n$ matrix $\mathbf{I} - \lambda_0 \mathbf{W}$. However, as a result it is not fully efficient. Lee (2003) introduces a best 2SLS estimator. This estimator uses the first expression for Ey in Eq. (80.32) in forming best instruments for \mathbf{Wy} . It is best in the sense that its asymptotic variance-covariance matrix is smallest among the class of GMM estimators based on linear moment conditions. Kelejian et al. (2004) introduce an alternative best 2SLS estimator (with identical asymptotic properties). For computational ease, this estimator uses a series approximation for the second expression for Ey in Eq. (80.32) when forming best instruments for \mathbf{Wy} . All of the above S2SLS estimators break down if $\beta_0 = 0$, i.e., if there are no exogenous variables in the model. This is not the case with the ML estimator. As a consequence, one would expect the ML estimator, given it is computable, to increasingly outperform the above S2SLS estimators as the variation in the disturbances increases relative to the variation in the regressors. However, Das et al. (2003) provide Monte Carlo results which suggest that the loss of efficiency of 2SLS-type estimators relative to ML estimation is modest for a wide range of specifications.

The above papers establish consistency of the GMM estimator for ρ_0 , but do not derive its limiting distribution. Drukker et al. (2011) derive the joint limiting distribution for two-step GMM estimators for δ_0 and ρ_0 .

Fingleton (2008) formulates moment conditions and GMM estimators for the case where the disturbance process is an MA rather than an AR process.

Lee (2007) considers an SARAR(1,0) model, i.e., model (80.22) with $\rho_0 = 0$. He suggests augmenting the usual linear moment conditions by quadratic moment conditions and derives the best quadratic moment condition. This best quadratic moment condition involves the inverse of $\mathbf{I} - \lambda_0 \mathbf{W}$. Lee shows that the corresponding best GMM estimator may have the same asymptotic distribution as

⁶For an incomplete list of empirical work see, e.g., Kelejian and Prucha (2010).

the ML estimator under normality. Also, the estimator does not break down if there are no explanatory exogenous variables. Liu et al. (2010) and Lee and Liu (2010) extend the results to one-step GMM estimators of an SARAR(1,1) and SARAR(p,q) model, respectively.

All of the above literature assumes that the basic innovations are homoskedastic. Kelejian and Prucha (2010) and Arraiz et al. (2010) consider two-step GMM estimation of an SARAR(1,1) model under the assumption that the innovations are heteroskedastic of unknown form. Badinger and Egger (2011) extend the approach the case of an SARAR(p,q) model. Lin and Lee (2010) consider one-step GMM estimation of an SARAR(1,0) with unknown heteroskedasticity, employing both linear and quadratic moment conditions.

Extensions of Cliff-Ord-type models to random and fixed effects panel data have been an important focus of recent research. Considered estimation methodologies have been GMM, quasi-ML, and Bayesian Markov Chain Monte Carlo methods.⁷ The literature on GMM estimation for panel data includes Kapoor et al. (2007), Murt and Pfaffermayr (2011), and Yu et al. (2012). Liu and Lee (2010) discuss GMM estimation (as well as other approaches) of a Cliff-Ord-type social interaction model. (See Elhorst, ► Chap. 82, “Spatial Panel Models”.)

Kelejian and Prucha (2007) and Drukker et al. (2011) discuss GMM estimation for Cliff-Ord-type single equation models with additional outside endogenous variables. Kelejian and Prucha (2004) consider a Cliff-Ord-type simultaneous equation system and discuss both limited and full information GMM estimators.

Pinkse et al. (2002) consider a semiparametric GMM approach, which allows for the spatial weights to be modeled as unknown functions of some distance measure. We note that if we are willing to assume that the weights can be expressed as, say, a finite polynomial in distance, then the substituted model will be of the form of an SARAR(p,q) model.

80.3.4 Exemplary GMM Estimators

In the following we give an illustrative result for the limiting distribution of GMM estimators for the SARAR(1,1) model (80.22). As remarked, for two-step GMM estimation, the limiting distribution of the GMM estimator for ρ_0 will depend on the estimator for δ_0 used in constructing estimated residuals. Our illustrative example will focus on the two-step GMM estimators considered in Kelejian and Prucha (1998, 1999), which can be viewed as a special case of the GMM estimators considered in Sects. 80.3.1 and 80.3.2, with $S_\rho = 2$ and

⁷Quasi-ML and Bayesian MCMC methods are not covered by this review. For recent papers employing those methods within the context of dynamic panel data models, see, e.g., Yu et al. (2008) and Parent and LeSage (2012), respectively. There is also an important literature on testing for spatial dependence in a panel context, which is not part of this review. For a partial review of this literature see, e.g., Baltagi (2011).

$\mathbf{A}_1 = v[\mathbf{M}'\mathbf{M} - n^{-1}\text{tr}(\mathbf{M}'\mathbf{M})\mathbf{I}]$ with $v = 1/\left[1 + [n^{-1}\text{tr}(\mathbf{M}'\mathbf{M})]^2\right]$ and $\mathbf{A}_2 = \mathbf{M}$. The discussion below assumes that the assumptions maintained in that paper hold, including that the innovations ε_i are i.i.d. $(0, \sigma^2)$.

We next describe specific steps in computing the GMM estimators.

Step 1a: 2SLS Estimator

In the first step, estimate δ by 2SLS from the untransformed model (80.22), using the instrument matrix \mathbf{H} as discussed in Sect. 80.3.2. The 2SLS estimator, say $\tilde{\delta}$, is then given by $\tilde{\delta} = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{y}$, where $\tilde{\mathbf{Z}} = \mathbf{P}_{\mathbf{H}}\mathbf{Z}$ with $\mathbf{P}_{\mathbf{H}} = \mathbf{H}(\mathbf{H}'\mathbf{H})^{-1}\mathbf{H}'$.

Step 1b: Initial GMM Estimator of ρ Based on 2SLS Residuals

Let $\tilde{\mathbf{u}} = \mathbf{u}(\tilde{\delta}) = \mathbf{y} - \mathbf{Z}\tilde{\delta}$ denote the 2SLS residuals. Consider the following sample moments based on estimated 2SLS residuals:

$$\mathbf{q}_n^\rho(\rho, \tilde{\delta}) = n^{-1} \begin{bmatrix} \tilde{\mathbf{u}}'(\mathbf{I} - \rho\mathbf{M}')\mathbf{A}_1(\mathbf{I} - \rho\mathbf{M})\tilde{\mathbf{u}} \\ \tilde{\mathbf{u}}'(\mathbf{I} - \rho\mathbf{M}')\mathbf{A}_2(\mathbf{I} - \rho\mathbf{M})\tilde{\mathbf{u}} \end{bmatrix} \quad (80.37)$$

The initial GMM estimator for ρ is then defined as

$$\tilde{\rho} = \arg \min_{\rho} \left\{ \mathbf{q}_n^\rho(\rho, \tilde{\delta})' \mathbf{q}_n^\rho(\rho, \tilde{\delta}) \right\}$$

Clearly $\tilde{\rho}$ is a special case of the class of estimators considered in Eq. (80.31) with $\mathbf{Y}_n^{\rho\rho} = \mathbf{I}$.

Step 2a: FGS2SLS Estimator

In the second step, reestimate δ by FGS2SLS, as discussed in Sect. 80.3.2. The FGS2SLS estimator is defined as the 2SLS estimator of the Cochrane-Orcutt transformed model (80.24) with the parameter ρ_0 replaced by $\tilde{\rho}$ computed in Step 1b. The FGS2SLS estimator is given by $\hat{\delta} = [\hat{\mathbf{Z}}_*(\tilde{\rho})'\mathbf{Z}_*(\tilde{\rho})]^{-1}\hat{\mathbf{Z}}_*(\tilde{\rho})'\mathbf{y}_*(\tilde{\rho})$ where $\hat{\mathbf{Z}}_*(\tilde{\rho}) = \mathbf{P}_{\mathbf{H}}\mathbf{Z}_*(\tilde{\rho})$.

Step 2b: Efficient GMM Estimator of ρ Based on FGS2SLS Residuals

Let $\hat{\mathbf{u}} = \mathbf{y} - \mathbf{Z}\hat{\delta}$ denote the FGS2SLS residuals, and let $\mathbf{q}_n^\rho(\rho, \hat{\delta})$ be defined as in Eq. (80.5) with $\tilde{\mathbf{u}}$ replaced by $\hat{\mathbf{u}}$. By Drukker et al. (2011), the corresponding efficient GMM estimator for ρ_0 based on FGS2SLS residuals is then given by

$$\hat{\rho} = \arg \min_{\rho} \left[\mathbf{q}_n^\rho(\rho, \hat{\delta})' (\hat{\Psi}_n^{\rho\rho})^{-1} \mathbf{q}_n^\rho(\rho, \hat{\delta}) \right]$$

where $\hat{\Psi}_n^{\rho\rho} = (\hat{\psi}_{rs}^{\rho\rho})_{r,s=1,2}$ is an estimator of the variance-covariance matrix of the limiting distribution of the normalized sample moments $n^{1/2}\mathbf{q}_n^\rho(\rho, \hat{\delta})$. In particular we have⁸

⁸In the following $\text{vec}_D(\mathbf{A})$ refers to the column vector containing the diagonal elements of the matrix \mathbf{A} .

$$\begin{aligned}\widehat{\psi}_{rs}^{\rho\rho} &= \widehat{\sigma}^4 (2n)^{-1} \text{tr}[(\mathbf{A}_r + \mathbf{A}'_r)(\mathbf{A}_s + \mathbf{A}'_s)] \\ &\quad + \widehat{\sigma}^2 n^{-1} \widehat{\mathbf{a}}'_r \widehat{\mathbf{a}}_s \\ &\quad + n^{-1} \left(\widehat{\mu}^{(4)} - 3 \widehat{\sigma}^4 \right) \text{vec}_D(\mathbf{A}_r)' \text{vec}_D(\mathbf{A}_s) \\ &\quad + n^{-1} \widehat{\mu}^{(3)} [\widehat{\mathbf{a}}'_r \text{vec}_D(\mathbf{A}_s) + \widehat{\mathbf{a}}_s \text{vec}_D(\mathbf{A}_r)]\end{aligned}$$

where $\widehat{\mathbf{a}}_r = \mathbf{H} \widehat{\mathbf{P}}^* \widehat{\alpha}_r$ and

$$\begin{aligned}\widehat{\mathbf{P}}^* &= (n^{-1} \mathbf{H}' \mathbf{H})^{-1} (n^{-1} \mathbf{H}' \mathbf{Z}_*(\widetilde{\rho})) \\ &\quad \times \left[(n^{-1} \mathbf{Z}'_*(\widetilde{\rho}) \mathbf{H}) (n^{-1} \mathbf{H}' \mathbf{H})^{-1} (n^{-1} \mathbf{H}' \mathbf{Z}_*(\widetilde{\rho})) \right]^{-1} \\ \widehat{\alpha}_r &= -n^{-1} \mathbf{Z}'_*(\widetilde{\rho}) (\mathbf{A}_r + \mathbf{A}'_r) \widehat{\varepsilon},\end{aligned}$$

and $\widehat{\sigma}_n^2$, $\widehat{\mu}_n^{(3)}$, and $\widehat{\mu}_n^{(4)}$ are standard sample estimators of σ^2 , $\mu^{(3)} = E\varepsilon_i^3$, and $\mu^{(4)} = E\varepsilon_i^4$ based on $\widehat{\varepsilon} = (\mathbf{I} - \widetilde{\rho} \mathbf{M}) \widehat{\mathbf{u}}$.

The derivation of the limiting distribution of $n^{1/2} \mathbf{q}^\rho(\rho, \widehat{\delta})$ used the CLT for linear quadratic forms given as Theorem 1. Observe that $\widehat{\alpha}_r$ is an estimator for $\alpha_r = -n^{-1} E \mathbf{Z}'_*(\rho) (\mathbf{A}_r + \mathbf{A}'_r) \varepsilon$. If the model does not contain a spatial lag in \mathbf{y} , i.e., if $\mathbf{Z} = \mathbf{X}$, then $\alpha_r = 0$ and we can then take $\widehat{\alpha}_r = 0$.

Based on Drukker et al. (2011), we now have the following result for the joint-asymptotic distribution of the final stage estimators $\widehat{\delta}$ and $\widehat{\rho}$:

$$\begin{bmatrix} \widehat{\delta} \\ \widehat{\rho} \end{bmatrix} \stackrel{\text{d}}{\sim} N \left(\begin{bmatrix} \delta_0 \\ \rho_0 \end{bmatrix}, n^{-1} \begin{bmatrix} \widehat{\Omega}^{\delta\delta} & \widehat{\Omega}^{\delta\rho} \\ \widehat{\Omega}^{\delta\rho'} & \widehat{\Omega}^{\rho\rho} \end{bmatrix} \right)$$

where

$$\begin{aligned}\widehat{\Omega}^{\delta\delta} &= \widehat{\mathbf{P}}^{*\prime} \widehat{\Psi}^{\delta\delta} \widehat{\mathbf{P}}^* \\ \widehat{\Omega}^{\delta\rho} &= \widehat{\mathbf{P}}^{*\prime} \widehat{\Psi}^{\delta\rho} \left(\widehat{\Psi}^{\rho\rho} \right)^{-1} \widehat{\mathbf{J}} \left[\widehat{\mathbf{J}}' \left(\widehat{\Psi}^{\rho\rho} \right)^{-1} \widehat{\mathbf{J}} \right]^{-1} \\ \widehat{\Omega}^{\rho\rho} &= \left[\widehat{\mathbf{J}}' \left(\widehat{\Psi}^{\rho\rho} \right)^{-1} \widehat{\mathbf{J}} \right]^{-1} \\ \widehat{\Psi}^{\delta\delta} &= \widehat{\sigma}^2 n^{-1} \mathbf{H}' \mathbf{H} \\ \widehat{\Psi}^{\delta\rho} &= \widehat{\sigma}^2 n^{-1} \mathbf{H}' [\widehat{\mathbf{a}}_1, \widehat{\mathbf{a}}_2] + \widehat{\mu}^{(3)} n^{-1} \mathbf{H}' [\text{vec}_D(\mathbf{A}_1), \text{vec}_D(\mathbf{A}_{2,n})]\end{aligned}$$

where $\widehat{\mathbf{P}}^*$, $\widehat{\mathbf{a}}_r$, and $\widehat{\Psi}^{\rho\rho}$ are as defined above, and

$$\widehat{\mathbf{J}} = n^{-1} \begin{bmatrix} 2 \widehat{\mathbf{u}}' \mathbf{M}' \mathbf{A}_1 \widehat{\mathbf{u}} & -\widehat{\mathbf{u}}' \mathbf{M}' \mathbf{A}_1 \mathbf{M} \widehat{\mathbf{u}} \\ 2 \widehat{\mathbf{u}}' \mathbf{M}' \mathbf{A}_2 \widehat{\mathbf{u}} & -\widehat{\mathbf{u}}' \mathbf{M}' \mathbf{A}_2 \mathbf{M} \widehat{\mathbf{u}} \end{bmatrix} \begin{bmatrix} 1 \\ 2\widehat{\rho} \end{bmatrix}$$

For interpretation, observe that $\widehat{\Omega}^{\delta\delta} = \widehat{\sigma}^2 [\widehat{\mathbf{Z}}_*(\widehat{\rho})' \widehat{\mathbf{Z}}_*(\widehat{\rho})]^{-1}$, i.e., the above expression for the estimator of variance-covariance matrix of the joint distribution of $\widehat{\delta}$ and $\widehat{\rho}$ delivers the usual estimator for the variance-covariance matrix of the FGS2SLS estimator as a special case.

The above joint-asymptotic-normality result allows a joint Wald test for the absence of spatial dependencies, i.e., a joint test of $H_0 : \lambda_0 = 0, \rho_0 = 0$.

80.4 GMM Estimation of Models with Spatial Mixing

Cliff-Ord-type models are linear simultaneous equation models where mixing of the data process is achieved through the assumption that the basic innovations are independently distributed combined with assumptions on the spatial weight matrices, such as that the row and column sums of their absolute elements are bounded.

In the time series literature, a widely used notion of dependence is α -mixing. This concept has been generalized to spatial processes (or random fields). In an important paper, Conley (1999) considered GMM estimators for stationary α -mixing spatial processes. Stationarity implies that the process has constant mean and variance and that the covariances only depend on distance (in a particular direction). Many economic processes are likely to exhibit some form of non-stationarity – e.g., housing prices may increase very much as we move toward the center of a city. Thus, relaxing the stationarity assumption seemed important.

One difficulty in developing a generalized theory of inference for spatial processes was a paucity of limit theorems (laws of large numbers, uniform laws of large numbers, and central limit theorems) which are sufficiently general. In light of this, Jenish and Prucha (2009) developed limit theorems for non-stationary α -mixing spatial processes, allowing also for the locations of observations to form a nonregular grid. Still, since α -mixing is not necessarily preserved under infinite lag formations, a further expansion of the theory to a class of spatial processes, which is closed under infinite lag formations, seemed desirable. To that effect Jenish and Prucha (2012) extended the notion of near-epoch dependence from the time series literature to spatial processes. They then developed limit theorems for possibly non-stationary spatial processes which are near-epoch dependent on an α -mixing process and gave results concerning the consistency and asymptotic normality of GMM estimators for this generalized class of processes.

In a recent publication, Robinson and Thawornkaiwong (2012) consider a partially linear regression model. They define a semiparametric instrumental variable estimator and give results on its asymptotic properties, allowing for spatial dependence in the regressors and disturbances.

80.5 Conclusion

Over the last two decades, significant strides have been made toward developing a formal methodology of inference for spatial models or, more generally, for cross-

sectional interaction models. GMM estimation has been an important part of this literature. As usual, empirical work often confronts us with more challenging realities than what can be handled by existing methodologies of inference, and much more work is needed.

Acknowledgments I would like to thank James LeSage and Pablo Salinas Macario for their helpful comments on this chapter.

References

- Anselin L (1988) Spatial econometrics: methods and models. Kluwer, Boston
- Anselin L (2010) Thirty years of spatial econometrics. *Pap Reg Sci* 89(1):3–25
- Arbia G (2006) Spatial econometrics, statistical foundations and applications to regional convergence. Springer, New York
- Arraiz I, Drukker DM, Kelejian HH, Prucha IR (2010) A spatial Cliff-Ord-type model with heteroskedastic innovations: small and large sample results. *J Reg Sci* 50(2):592–614
- Badinger H, Egger P (2011) Estimation of higher-order spatial autoregressive cross-section models with heteroscedastic disturbances. *Pap Reg Sci* 90(1):213–235
- Baltagi BH (2011) Spatial panels. In: Ullah A, Giles DEA (eds) *The handbook of empirical economics and finance*. Chapman and Hall, Boca Raton, pp 435–454
- Cliff A, Ord J (1973) Spatial autocorrelation. Pion, London
- Cliff A, Ord J (1981) Spatial processes, models and applications. Pion, London
- Conley T (1999) GMM estimation with cross sectional dependence. *J Econ* 92:1–45
- Cressie N (1993) Statistics of spatial data. Wiley, New York
- Das D, Kelejian HH, Prucha IR (2003) Small sample properties of estimators of spatial autoregressive models with autoregressive disturbances. *Pap Reg Sci* 82(1):1–26
- Drukker DM, Egger P, Prucha IR (2011) On two-step estimation of a spatial autoregressive model with autoregressive disturbances and endogenous regressors. *Econ Rev* (forthcoming)
- Fingleton B (2008) A generalized method of moments of moments estimator for a spatial model with moving average errors, with application to real estate prices. *Empirical Economics* 34:35–57
- Haining R (2003) Spatial data analysis, theory and practice. Cambridge University Press, Cambridge
- Jenish N, Prucha IR (2009) Central limit theorems and uniform laws of large numbers for arrays of random fields. *J Econ* 150(1):86–98
- Jenish N, Prucha IR (2012) On spatial processes and asymptotic inference under near-epoch dependence. Department of Economics University of Maryland, Mimeo
- Kapoor M, Kelejian HH, Prucha IR (2007) Panel data models with spatially correlated error components. *J Econ* 140(1):97–130
- Kelejian HH, Prucha IR (1998) A generalized spatial two-stage least squares procedure for estimating a spatial autoregressive model with autoregressive disturbances. *J Real Estate Fin Econ* 17(1):99–121
- Kelejian HH, Prucha IR (1999) A generalized moments estimator for the autoregressive parameter in a spatial model. *Int Econ Rev* 40(2):509–533
- Kelejian HH, Prucha IR (2001) On the asymptotic distribution of the Moran I test statistic with applications. *J Econ* 104(2):219–257
- Kelejian HH, Prucha IR (2004) Estimation of simultaneous systems of spatially interrelated cross sectional equations. *J Econ* 118(1–2):27–50
- Kelejian HH, Prucha IR (2007) HAC estimation in a spatial framework. *J Econ* 140(1):131–154

- Kelejian HH, Prucha IR (2010) Specification and estimation of spatial autoregressive models with autoregressive and heteroskedastic disturbances. *J Econ* 157(1):53–67
- Kelejian HH, Prucha IR, Yuzefovich E (2004) Instrumental variable estimation of a spatial autoregressive model with autoregressive disturbances: large and small sample results. In: LeSage JP, Pace PR (eds) *Advances in econometrics: spatial and spatiotemporal econometrics*. Elsevier, New York, pp 163–198
- Lee L-F (2003) Best spatial two-stage least squares estimators for a spatial autoregressive model with autoregressive disturbances. *Econ Rev* 22(4):307–335
- Lee L-F (2004) Asymptotic distributions of maximum likelihood estimators for spatial autoregressive models. *Econometrica* 72(6):1899–1925
- Lee L-F (2007) GMM and 2SLS estimation of mixed regressive, spatial autoregressive models. *J Econ* 137(2):489–514
- Lee L-F, Liu X (2010) Efficient GMM estimation of higher order spatial autoregressive models with autoregressive disturbances. *Econ Theory* 26(1):187–230
- LeSage JP, Pace RK (2009) *Introduction to spatial econometrics*. CRC Press/Taylor and Francis, Boca Raton
- Lin X, Lee L-F (2010) GMM estimation of spatial autoregressive models with unknown heteroskedasticity. *J Econ* 157(1):34–52
- Liu X, Lee L-F (2010) GMM estimation of social interaction models with centrality. *J Econ* 159(1):99–115
- Liu X, Lee L-F, Bollinger CR (2010) An efficient GMM estimator of spatial autoregressive models. *J Econ* 159(2):303–319
- Mutl J, Pfaffermayr M (2011) The Hausman test in a Cliff and Ord panel model. *Econ J* 14(1):48–76
- Paelink JHP, Klaassen LH (1979) *Spatial econometrics*. Saxon House, Farnborough
- Parent O, LeSage JP (2012) Spatial dynamic panel data models with random effects. *Reg Sci Urban Econ* 42(4):727–738
- Pinkse J, Slade ME, Brett C (2002) Spatial price competition: a semiparametric approach. *Econometrica* 70(3):1111–1153
- Robinson PM, Thawornkaiwong S (2012) Statistical inference on regression with spatial dependence. *J Econ* 167(2):521–542
- Tobler W (1970) A computer movie simulating urban growth in the Detroit region. *Econ Geogr* 46(2):234–240
- Whittle P (1954) On stationary processes in the plane. *Biometrika* 41(3/4):434–449
- Yu J, de Jong R, Lee L-F (2008) Quasi-maximum likelihood estimators for spatial dynamic panel data with fixed effects when both n and T are large. *J Econ* 146(1):118–134
- Yu J, de Jong R, Lee L-F (2012) Estimation for spatial dynamic panel data with fixed effects: the case of spatial cointegration. *J Econ* 167(1):16–37

Xiaokun (Cara) Wang

Contents

81.1	Introduction	1620
81.2	Limited and Censored Variable Models	1621
81.2.1	Models for Discrete Responses	1621
81.2.2	Models for Censored and Truncated Data	1623
81.3	Models Incorporating Spatial Effects	1624
81.3.1	Geographically Weighted Regression	1624
81.3.2	Spatial Filtering	1625
81.3.3	Spatial Regression	1626
81.4	Estimation Approaches	1627
81.4.1	Maximum Simulated Likelihood Estimation (MSLE)	1628
81.4.2	Composite Marginal Likelihood	1629
81.4.3	Bayesian Approach	1631
81.5	Conclusions	1633
	References	1634

Abstract

In regional science, many attributes, either social or natural, can be categorical. For example, choices of travel mode, presidential election outcomes, or quality of life can all be measured (and/or coded) as discrete responses, dependent on various influential factors. Some attributes, although continuous, are subject to truncation or censoring. For example, household income, when reported, tends to be censored, and only boundary values of a range are obtained. Such categorical and censored variables can be analyzed using econometric models that are established based on the concept of “unobserved/latent dependent variable.”

X. Wang

Department of Civil and Environmental Engineering, Rensselaer Polytechnic Institute,
Troy, NY, USA
e-mail: wangx18@rpi.edu

The previous examples also share another common feature: when data is collected in a spatial setting, they are all inevitably influenced by spatial effects, either spatial variation or spatial interaction. In contrast to panel data or time-series data, such variation or dependencies are two-dimensional, making it even more complicated. The need for investigating such limited and censored variables in a spatial context compels the quest for rigorous statistical methods.

This chapter introduces existing methods that are developed to analyze limited and censored dependent variables while considering the spatial effects. Different model specifications are discussed, with an emphasis on discrete response models and censored data models. Different types of spatial effects and corresponding ways to address them are then discussed. In general, when the spatial variation is of major concern, geographically weighted regression is preferred. When the spatial dependency is the primary interest, spatial filtering and spatial regression should be chosen. Techniques popularly used to estimate spatial limited variable models, including maximum simulated likelihood estimation, composite marginal likelihood estimation, and Bayesian approach, are also introduced and briefly compared.

81.1 Introduction

In studies of social behaviors and human activities, many attributes involve categorical, truncated, or censored responses in a spatial context. For example, choices of travel mode, choices of occupation, and presidential election outcomes can all be measured (and/or coded) as discrete responses, dependent on various influential factors. Household income and pavement surface deterioration levels, when reported, tend to be censored or truncated. Such categorical and/or censored variables can be analyzed using the limited dependent variable models. The previous examples also share a common feature: they all exhibit some type of spatial effects, either spatial variation or some degree of spatial dependence. For example, in studies of ecology, wealth, vehicle crashes, and epidemics, it is known that the data generation processes often vary over space. Another example is that even after controlling for household attributes, choice of travel mode is still expected to exhibit positive spatial correlations. Such correlation patterns can be partly explained by proximity because, in reality, there are always influential factors that cannot be controlled (e.g., pedestrian friendliness of all neighborhoods). The sign and magnitude of such dependence tend to vary rather gradually over space. Most likely, correlation diminishes with increases in distance between any two observation units. And in a spatial context, in contrast to time-series data, such dependencies are two-dimensional – which adds complexity. The widespread nature of such phenomena and the need for understanding these behaviors compel the quest for rigorous statistical methods for analysis of such data.

However, the handling of limited and censored dependent variables already involves specification and estimation of nonlinear models. Considering spatial

effects implies that the models have to further account for two-dimensional dependence structures across a large number of observations, leading to manipulation of high-dimensional multivariate distributions and large matrices. In recent years, many studies have attempted to enhance the behavioral consistency of model specification and the efficiency of estimation. This chapter will introduce the related methods developed to date. The following sections will first introduce conventional econometric models used for limited and censored dependent variables, followed by a discussion of spatial models, that is, how the consideration of spatial effects can be incorporated. Estimation techniques, which are critical for the application of these models, will be discussed in the end.

81.2 Limited and Censored Variable Models

Models for limited and censored variables are an important subarea of econometrics. This section will explain the specification of these models in two main categories: those for discrete responses and those for censored/truncated data.

81.2.1 Models for Discrete Responses

Models for discrete responses are often used to model choices among sets of alternatives rather than a continuous response (Greene 2002). Such models play an important role in scientific studies, both social and natural. The specification of discrete response models tends to require specific assumptions on the error term distribution. Two commonly used specifications are probit and logit models. The most basic form is a binary response, where the value of the dependent variable is either 0 or 1, indicating no or yes:

$$\vec{y_i} = X_i' \beta + \varepsilon_i \text{ and } y_i = \begin{cases} 1 & \text{if } \vec{y_i} > 0 \\ 0 & \text{if } \vec{y_i} \leq 0 \end{cases} \quad i = 1, 2, \dots, N \quad (81.1)$$

where i indexes observations ($i = 1, 2, \dots, N$), $\vec{y_i}$ is a latent (unobserved) dependent variable for individual i , and y_i is the observed dependent variable. X_i is a $Q \times 1$ vector of explanatory variables, and β is the set of corresponding parameters. ε_i stands for unobservable factors for observation i and is assumed to follow an identically and independently distributed (iid) standard normal distribution for a probit model or Gumbel distribution for a logit model. In other words, the actual response that is observed is a nonlinear function of latent response, which can be expressed as a linear function of explanatory variables. With this model setting, it is straightforward to show that

$$\begin{aligned} Pr(y_i = 1|X_i) &= Pr(\vec{y_i} > 0|X_i) = Pr(\varepsilon_i > -X_i' \beta|X_i) = F(X_i' \beta) \\ Pr(y_i = 0|X_i) &= 1 - F(X_i' \beta) \end{aligned} \quad (81.2)$$

where $F(\cdot)$ is the cumulative distribution function (CDF) of error term ε_i . The log-likelihood is thus

$$\ln L = \sum_{i=1}^N \{y_i \ln F(X_i' \beta) + (1 - y_i) \ln[1 - F(X_i' \beta)]\} \quad (81.3)$$

Of course, in many circumstances, the number of alternatives is more than two. If the data are ordered and each unit makes a choice from among the S alternatives, the model specification can be naturally extended from the binary choice setting, with a set of threshold parameters to distinguish different levels of response (alternatives):

$$y_i = k \text{ if } \gamma_k < y_i^\rightarrow < \gamma_{k+1} \quad k = 1, 2, \dots, S \quad (81.4)$$

To some extent, the observed variable can be considered as a censored form of the latent variable: The latent variable y_i^\rightarrow varies continuously, but the observed response is censored by unknown boundaries $\gamma_1 < \gamma_2 < \dots < \gamma_{S+1}$, leading to one of the integer responses $1, 2, \dots, S$. The probability for each outcome is

$$\begin{aligned} Pr(y_i = 1|X_i) &= F(\gamma_2 - X_i' \beta) - F(\gamma_1 - X_i' \beta) \\ Pr(y_i = 2|X_i) &= F(\gamma_3 - X_i' \beta) - F(\gamma_2 - X_i' \beta) \\ &\vdots \\ Pr(y_i = S|X_i) &= F(\gamma_{S+1} - X_i' \beta) - F(\gamma_S - X_i' \beta) \end{aligned} \quad (81.5)$$

where $F(\cdot)$ is still the CDF of error term ε_i , a standard normal distribution in probit model and logistic in logit model.

When the data is multinomial and unordered, a common model specification is established based on the utility maximization theory introduced by McFadden (1980). In this framework, the alternative offering the maximum utility is chosen. If U_{ik} indicates utility for individual i to select alternative k ($k = 1, 2, \dots, S$), the observed dependent variable for observation i , y_i , will take value m if and only if U_{im} is the maximum utility among all alternatives (i.e., the most attractive option). Furthermore, one response alternative is often chosen as the “base” since preference is always a relative term. If the last alternative S is used as the base, the latent utility difference can be expressed as

$$y_{ik}^\rightarrow = U_{ik} - U_{is} \quad k = 1, 2, \dots, S-1 \quad (81.6)$$

Similarly, the latent utility difference is influenced by many factors, so

$$y_{ik}^\rightarrow = X_{ik}' \beta + \varepsilon_{ik} \text{ and } y_i = m \text{ if } y_{im}^\rightarrow > 0 \text{ and } y_{im}^\rightarrow \geq y_{ik}^\rightarrow \quad k = 1, 2, \dots, S-1 \quad (81.7)$$

X'_{ik} is a $1 \times Q$ vector indicating the differences of explanatory variable values between alternative k and base alternative S . Subscript k implies that X'_{ik} can be alternative specific (such as cost of different modes in the analysis of travel mode choice). Conventional models used for analyzing unordered categorical data are multinomial logit or multinomial probit models. When the iid assumption is potentially violated, there are other derived forms to deal with the correlated errors, for example, the nested logit model which requires prespecified error correlation structure and the random parameter (mixed) logit (probit) models that assume parameters follow random distributions.

81.2.2 Models for Censored and Truncated Data

A sample is considered “truncated” when it is only a subset of a larger population. For example, when studies on expenditures are based on observations with positive expenditures, those with no expenditures are “truncated.” A similar and more common problem is “censoring,” meaning that rather than observing the exact value, only the boundary value of a range is observed. With the previous example, if expenditure over \$10,000 is coded as \$10,000, it is censored from above. Censored and truncated data are not representative of the population, and estimators that ignore this problem will be inconsistent, leading to incorrect marginal effects (Greene 2002). To some extent, truncated and censored regressions are similar to binary and ordered-response models. A latent dependent variable is posited taking the form $y_i^\rightarrow = X'_i\beta + \varepsilon_i$, then for truncated (from below at 0) regression, the observed dependent variable will be

$$y_i = y_i^\rightarrow \quad \text{if } y_i^\rightarrow > 0 \quad (81.8)$$

If ε_i follows a normal distribution with zero mean and variance σ^2 , it can be shown that the conditional mean of the dependent variable is

$$\begin{aligned} E(y_i|y_i^\rightarrow > 0) &= E(X'_i\beta + \varepsilon_i|X'_i\beta + \varepsilon_i > 0) = X'_i\beta + E(\varepsilon_i|\varepsilon_i > -X'_i\beta) \\ &= X'_i\beta + \sigma \frac{\phi\left(\frac{X'_i\beta}{\sigma}\right)}{\Phi\left(\frac{X'_i\beta}{\sigma}\right)} \end{aligned} \quad (81.9)$$

where $\phi(\cdot)$ and $\Phi(\cdot)$ are the probability density function (PDF) and cumulative distribution function (CDF) of a standard normal distribution, respectively. In other words, because of the truncation, the mean of the dependent variable is no longer a linear function of X'_i . For censored (also from below at 0) regression, or Tobit model, the observed dependent variable will be

$$y_i = \begin{cases} y_i^\rightarrow & \text{if } y_i^\rightarrow > 0 \\ 0 & \text{if } y_i^\rightarrow \leq 0 \end{cases} \quad (81.10)$$

The censored conditional mean is thus $E(y_i|X_i) = Pr(y_i^\rightarrow \leq 0)0 + Pr(y_i^\rightarrow > 0)E(y_i|y_i^\rightarrow > 0)$, and with results for the truncated mean, it can be shown that

$$\begin{aligned} E(y_i|X_i) &= 0 + \Phi\left(\frac{X_i'\beta}{\sigma}\right) \left(X_i'\beta + \sigma \frac{\phi\left(\frac{X_i'\beta}{\sigma}\right)}{\Phi\left(\frac{X_i'\beta}{\sigma}\right)} \right) \\ &= \Phi\left(\frac{X_i'\beta}{\sigma}\right) X_i'\beta + \sigma \phi\left(\frac{X_i'\beta}{\sigma}\right) \end{aligned} \quad (81.11)$$

The above derivation can be easily extended to other truncation/censoring threshold values and the double truncation/censoring situation. As these conditional means are nonlinear, ordinary least square (OLS) will no longer yield consistent estimates of β . A rich body of literature related to generalization of censored models can be found in discussions of sample selection models and treatment effects models. It should be mentioned that, to some extent, duration models and count data models can be also considered as members of the limited dependent variable model family. Although their specification forms differ significantly from the previously discussed models, the notion behind their model specification is similar. If further investigation of these models is of interest, readers are referred to work by Greene (2002) which provides more in-depth discussion of these models.

81.3 Models Incorporating Spatial Effects

Methods for dealing with spatial effects in limited and censored dependent variable models can be categorized into the three types. The first method is geographically weighted regression (GWR), where the consideration is mainly about the spatial variation of behavioral parameters. The second method, spatial filtering, has been applied more broadly. This approach essentially attempts to “filter” spatial effect by explicitly controlling for variables that represent the spatial dependency. The third method is spatial regression, which incorporates spatial effects explicitly in the model specification, either through the spatial autocorrelation of (dependent and independent) variables, the autocorrelation of error terms, or both. This section introduces the basic concepts of these methods and discusses how these approaches can be integrated into the limited and censored dependent variable models.

81.3.1 Geographically Weighted Regression

GWR is established based on the assumption that relationships between variables vary from location to location; therefore, parameters should exhibit significant spatial variation. The flexible specification of GWR can be used to examine the stability and robustness of parameter estimates over space. The formulation of

GWR is fairly straightforward: Instead of a global regression that implies one data generation process dominates the whole population, the model is now localized by allowing for one unique data generation process per observation. Taking a binary model, for example, instead of a universal regression model for the latent variable $y_i^\rightarrow = X_i' \beta + \varepsilon_i$, the parameters are allowed to be local, that is,

$$y_i = X_i' \beta(u_i, v_i) + \varepsilon_i \text{ and } y_i = \begin{cases} 1 & \text{if } y_i^\rightarrow > 0 \\ 0 & \text{if } y_i^\rightarrow \leq 0 \end{cases} \quad i = 1, 2, \dots, N \quad (81.12)$$

where (u_i, v_i) indicate the coordinates of observation i and $\beta(u_i, v_i)$ is a continuous function of the map coordinates. The key advantage of GWR is that it explicitly allows for local spatial effects in relatively standard regression models (Fotheringham 2003). Still using the binary choice model as the example, with GWR, the log-likelihood for the j^{th} observation will be

$$\ln L_j = \sum_{i=1}^N \{ w_{ji} \{ y_i \ln F(X_i' \beta_j) + (1 - y_i) \ln [1 - F(X_i' \beta_j)] \} \} \quad (81.13)$$

where w_{ji} is the weight for the i^{th} data point with respect to the j^{th} regression point, normally higher for data points close to the j^{th} regression point and decays over distance. Comparing this expression with Eq. (81.3), it can be observed that the key differences are that each observation now has its own parameter values and that the regression is influenced more by data points nearby. Selecting weight function, or kernel, to define w_{ji} is often the most challenging step of GWR. The main considerations are the representation of point proximity and selection of bandwidth distance (or the cutoff distance over which the data points no longer influence the regression). Fotheringham (2003) describes a variety of weight specification alternatives, with Gaussian weights and their bi-square variation as the most commonly used options. The integration of GWR into limited dependent variable models is straightforward and has been applied in many studies. LeSage (1999) provided MATLAB code for estimating binary logit and probit GWR models, using crime data as an illustration. Atkinson et al. (2003) used a binary logit GWR model to identify relationships between geomorphological controls and riverbank erosion. McMillan and McDonald (1999) extended the use of GWR into multinomial discrete response analysis by specifying a multinomial logit GWR model to analyze the influence of transportation access on land use in Chicago. Luo and Wei (2009) analyzed land-use conversion (from barren, crop/grassland, forest, and water uses to urban land use) via a multinomial logit GWR model.

81.3.2 Spatial Filtering

When spatial interaction, rather than spatial variation, is the major concern, the approaches used to address the spatial dependency include spatial filtering and

spatial regression. “Spatial filtering” has many different definitions in existing literature. The most unrestricted definition is to simply construct and control for some spatial variables so that conventional statistical models based on uncorrelated errors could still apply. For example, Dugundji and Walker (2005) considered spatial network independencies in their mixed logit model when studying mode choice behavior. In most recent studies, “spatial filtering” is referred as the semiparametric approaches that separate spatial dependencies by dividing original variables into filtered nonspatial variables and spatial variables. The division is often achieved using local spatial statistics such as distance-based eigenvector procedure (Dray et al. 2006), G-statistics-based approach (Getis 1995), and eigenfunction-based procedure (Griffith 2000). The local spatial statistics such as Getis’ G_i and Moran’s I were originally developed as diagnostics to disclose local spatial dependencies that are not indicated by their global counterparts. For example, Getis’ G_i is established based on an indicator $G_i(d)$, which is essentially a weighted average of observation values around observation i :

$$G_i(d) = \frac{\sum_j w_{ij} x_j}{\sum_j x_j}, \text{ all } j \neq i \quad (81.14)$$

where w_{ij} is the element of a row-standardized geographic connectivity matrix W and d indicates the distance or other predefined connectivity index. Getis (1995) then used the difference between observed value $G_i(d)$ and expected value $E(G_i)$, to separate spatial from nonspatial effects. In other words, observed variables were filtered as

$$\vec{x}_i = x_i \frac{E(G_i)}{G_i(d)} = \frac{x_i \left[\frac{\sum_j w_{ij}}{n-1} \right]}{G_i(d)} \quad (81.15)$$

After all variables (both dependent and explanatory variables) have been filtered by such procedure, and the spatial dependency is considered removed, the conventional models introduced previously can be used directly for data analysis. Other spatial filtering techniques use different approaches to filter the spatial effects, but they all rely on the construction and manipulation of a spatial weight matrix W , which is used to represent the spatial dependency structure. The key challenge for this approach is to choose proper regional weighting scheme, or the construction of W .

81.3.3 Spatial Regression

Similar to spatial filtering, spatial regression models also directly address the spatial dependencies by incorporating spatial effects in the model specification. The key difference between spatial regression and spatial filtering is that, although spatial

regression may also use spatially lagged explanatory and/or response variables, as in the spatial filter models, many of the variables are treated as endogenous. LeSage and Pace (2009) summarized the motivations of using spatial regression models for data analysis. There is a big family of spatial regression models, including spatial autoregressive (SAR), spatial moving average (SMA), spatial Durbin (SDM), and spatial error (SEM), with SAR and SEM as the most commonly used specifications. Many existing works (Anselin 2003; LeSage and Pace 2009) have provided extensive technical discussions on these models and the underlying spatial stochastic processes. In general, SAR and SEM are used for dependent variables and error terms, respectively. Both are used rather regularly by researchers, thanks to their flexibility and applicability. The former case is also called “spatial lag,” while the latter is often called “spatial error.” By using these two specifications, it is assumed that the spatial process follows a recursive pattern. In a linear model setting, SAR is expressed as

$$y = \rho Wy + X\beta + \varepsilon \text{ or } y = (I - \rho W)X\beta + (I - \rho W)\varepsilon \quad \varepsilon \sim N(0, \sigma^2 I_n) \quad (81.16)$$

where W still denotes the geographic connectivity matrix, ρ is the spatial coefficient, representing the magnitude of overall neighborhood influence, and I_n is an n by n identity matrix.

The SEM incorporates the spatial effects in error terms:

$$y = X\beta + u \quad u = \lambda u + \varepsilon \quad \varepsilon \sim N(0, \sigma^2 I_n) \quad (81.17)$$

where u is the vector of overall errors, λ is the spatial coefficient for error terms, indicating the contribution of neighboring observations on each other's uncertainty, and ε now indicates the part of error or uncertainty caused by each observation itself. These spatial processes can be easily applied in the context of limited and censored dependent variable models by incorporating them in the formulation of latent dependent variables. For example, a SAR probit model is simply

$$\vec{y} = (I - \rho W)X\beta + (I - \rho W)\varepsilon \quad \varepsilon \sim N(0, \sigma^2 I_n) \text{ and still } y = \begin{cases} 1 & \text{if } \vec{y} > 0 \\ 0 & \text{if } \vec{y} \leq 0 \end{cases} \quad (81.18)$$

Many studies have used spatial regression models to analyze limited and censored dependent variables. For example, Beron and Vijverberg (1999) specified probit models with both spatial errors and spatial lags. Smith and LeSage (2004) incorporated a regional effect in a probit model and used Bayesian techniques to analyze the 1996 presidential election results. LeSage and Pace (2009) discussed specification of a spatial autoregressive multinomial probit model. Wang and Kockelman (2009) developed a spatial ordered probit model with temporal correlation, and Wang et al. (2012) further extended the models to the analysis of multinomial, unordered responses.

81.4 Estimation Approaches

A common approach for estimating limited and censored dependent variable models is the maximum likelihood estimation (MLE) technique (see Pace, ► Chap. 78, “Maximum Likelihood Estimation”). When the models are further complicated by the consideration of spatial dependency, implying interdependence of observations, the joint distribution of the entire sample is no longer the product of their marginal distributions; hence, the log-likelihood is no longer additively separable one-dimensional probabilities. The calculation of high-dimensional distribution requires the manipulation of large matrices and a high-dimensional integral. The MLE approach thus becomes ineffective facing such heavy, sometimes impossible computational burdens. Alternative estimation approaches have been explored by researchers. For example, Pinkse and Slade (1998) used the generalized method of moments (GMM) to estimate a probit model with spatial error components (see Prucha, ► Chap. 80, “Instrumental Variables/Method of Moments Estimation” for details concerning estimation). Klier and McMillen (2008) used GMM to estimate a spatial logit model for analyzing the clustering of auto supplier plants in the USA. McMillen (1995) used simulated likelihood strategies to estimate spatial multinomial probit models. Vijverberg (1997) used recursive importance sampling (RIS) to approximate the n-dimensional log-likelihood in a spatial probit model. Bhat (2011) suggested a maximum approximate composite marginal likelihood (MACML) method, which essentially decomposes multidimensional autocorrelation into pairwise correlation. Smith and LeSage (2004), LeSage and Pace (2009), and Wang and Kockelman (2009) used Bayesian framework in their studies of spatial discrete response models. In general, the estimation approaches discussed in existing literature can be categorized into four types: maximum simulated likelihood estimation (MSLE), GMM, MACML, and Bayesian techniques. Among them, the use of GMM is relatively limited because it requires orthogonality, which cannot be conveniently derived in multiple-response models. This section will briefly introduce the other three estimation techniques that are used more dominantly in practice: MSLE, MACML, and Bayesian techniques.

81.4.1 Maximum Simulated Likelihood Estimation (MSLE)

The notion of MSLE is that, since the direct derivation of complex statistical models is impractical (e.g., when a likelihood function involves a multidimensional integral), a simulated likelihood that approximates the original likelihood is used instead. When approximated appropriately, the key model features in the original model are retained, but computational burden is alleviated. For example, in a SAR multinomial probit model where the probability involves multivariate normal cumulative density function

$$Pr(V < v) = \Phi(V_1 < v, V_2 < v_2, V_3 < v_3, \dots, V_n < v_n) \quad (81.19)$$

where V is a vector of joint events and v is the corresponding thresholds. $\Phi(\cdot)$ still indicates the CDF of a normal distribution. The joint probability can be decomposed into the product of conditional densities:

$$Pr(V < v) = \int_{-\infty}^{v_1} \int_{-\infty}^{v_{I-1}} \cdots \int_{-\infty}^{v_1} \phi(V_I | V_{i < I}) \phi(V_{I-1} | V_{i < I-1}) \cdots \phi(V_1) dV_1 \cdots dV_I \quad (81.20)$$

The above distribution involves an intractable integral. One approach used to simulate this joint probability is GHK (Geweke-Hajivassiliou-Keane) simulator, which is considered most effective among traditional techniques. The GHK method uses the Cholesky decomposition to generate recursive ordering. Furthermore, instead of using the conditional probabilities reliant on the random event $V_{i < I-1}$ and calculating the multidimensional integral, a set of realized discrete V values are used. Each time-specific V values are given, denoted as \tilde{V} , the product of these conditional probabilities could be easily calculated based on these specific realizations, leading to an approximation of the original likelihood:

$$\widetilde{Pr}(V < v) = \int_{-\infty}^{v_1} \phi(V_I | \tilde{V}_{i < I}) dV_I \int_{-\infty}^{v_{I-1}} \phi(V_{I-1} | \tilde{V}_{i < I-1}) dV_{I-1} \cdots \int_{-\infty}^{v_1} \phi(V_1) dV_1 \quad (81.21)$$

“Simulated likelihood” generates the \tilde{V} vector values for multiple times (denoted as R). As Train (2003) explains, the simulation bias and noise are inversely proportional to R . Besides increasing R , researchers have also found that use of quasi-random numbers may enhance the performance of simulation. For example, Train (2003) suggested Halton sequences and its related variants, and Wang and Kockelman (2008) assessed several quasi-random number generation techniques and recommended the use of scrambled Halton sequence based on its performance on spatial data.

Beron and Vijverberg (1999) used the GHK method, also called recursive importance sampling (RIS), when estimating spatial probit models. However, one limitation is that the computational time of GHK increases exponentially with sample size, making it infeasible to handle large sample size. In a recent study, Pace and LeSage (2011) recognized the fact that spatial dependency often decays fast and leads to sparse matrix, which can be exploited with a sparse GHK procedure to improve the efficiency of spatial probit models for large-sized samples.

81.4.2 Composite Marginal Likelihood

Rather recently, some researchers have proposed the use of simplified pseudo-likelihoods that shares similar notion with MSLE but is claimed to be more efficient. The composite marginal likelihood (CML) method is one such approach, building pseudo-likelihoods by compounding low-dimensional “marginal” probabilities. CML studies date back to the 1970s, when the method was referred to as a partial likelihood, pseudo-likelihood, or quasi-likelihood approach. The term “composite

likelihood” was first proposed in 1988, and the approach has been gaining popularity since 2004, as more and more time-series and spatial-data studies have confronted estimation barriers. The method seeks computational savings at the cost of some loss in estimator efficiency. Popularly adopted approaches include the following:

- (a) Univariate distributions, which ignore dependence across observations, so the likelihood is referred to as an independent likelihood (IL). The use of univariate probabilities ignores all correlations and thus cannot recover parameters representing error correlation (and spatial) effects, but it will still lead to consistent estimators for the primary variables.
- (b) Bivariate distributions, which result in a pairwise likelihood (PL) function, or the multiple of pairwise marginal likelihoods:

$$PL(\beta; y) = \prod_{i=1}^{N-1} \prod_{j=i+1}^N f(y_i, y_j; \beta)^{w_{ij}} \quad (81.22)$$

where $f(\cdot)$ indicates the pairwise marginal likelihoods, i, j are two among a total of N observations, and w_{ij} denotes the weight assigned to the corresponding pairwise likelihood, normally set to be 1. Similarly, one can reflect three-way dependencies, using observation triplets, or combine the IL and PL likelihoods. Varin (2008) has described CML properties in detail, including the application of CML in spatial econometric settings. Of course, the suitability of a CML approach depends on whether it is appropriate to replace a high-dimensional joint distribution with low-dimensional marginal distributions. In general, for spatial data, where there are long sequences of correlated response values, CML inference retains good properties “provided that the data may be seen as formed by pseudo independent subsets. This is the case of stationary time-series and spatial processes with good mixing properties as an autocorrelation function exponentially decaying to zero.” (Varin 2008) In other words, if the autocorrelation dies down quickly, so observations are only weakly correlated with those “far” from their location, CML estimators tend to be quite efficient. If, instead, the spatial autocorrelation is rather global and does not decay fast, the convergence of CML estimators may be slow or even fail. The CML approach is now being tested by researchers from a wide spectrum of disciplines, including gene mapping and population dynamics. Bhat (2011) further extended CML and developed the maximum approximate composite marginal likelihood (MACML) by integrating the multivariate standard normal cumulative distribution (MVNCDF) function approximation and the CML. The MVNCDF function approximation (i.e., the “MA” part of MACML) changes the following joint probability $Pr(V < v) = \Phi(V_1 < v_1, V_2 < v_2, V_3 < v_3, \dots, V_I < v_I)$ into the product of a bivariate marginal probability and a univariate conditional probability:

$$Pr(V < v) = Pr(V_1 < v_1, V_2 < v_2) \times \prod_{i=3}^I Pr(V_i < v_i | V_1 < v_1, V_2 < v_2, \dots, V_{i-1} < v_{i-1}) \quad (81.23)$$

where the meaning of V and v is the same as those in Eq. (81.19). The right-hand side univariate conditional probability can be further approximated using linear regression models because

$$\begin{aligned} & Pr(V_i < v_i | V_1 < v_1, V_2 < v_2, \dots, V_{i-1} < v_{i-1}) \\ &= E\left(\widetilde{I}_i | \widetilde{I}_1 = 1, \widetilde{I}_2 = 1, \widetilde{I}_3 = 1, \dots, \widetilde{I}_{i-1} = 1\right) \end{aligned} \quad (81.24)$$

where \widetilde{I}_i is a binary indicator corresponding to $V_i < v_i$. Essentially, the MA part handles the correlation within one event (observation), and the CML takes care of the correlation across events (observations). Bhat (2011) discussed MACML's application in spatial data sets. If the spatial dependency is global (extending across all Q observations), there will be $Q(Q-1)/2$ pairs of bivariate probabilities. Bhat (2011) claims that “... in a spatial case where dependency drops quickly with inter-observation distance, the pairs formed from the closest observations provide much more information than pairs that are very far away. ... Typically, in a spatial context, there appears to be an optimal distance for inclusion of observation pairs. This distance threshold may be set based on knowledge about the spatial process or based on testing the efficiency of estimators with varying values of the distance threshold. Using such a distance threshold effectively reduces the number of pairwise terms in the CML function.” Paleti and Bhat (2011) also compared the performance of CML to the conventional MSLE approach and concluded that the CML approach recovers the parameter well and is more efficient than MSLE. Bhat and coauthors then applied MACML in various settings involving spatial limited dependent variable models: Ferdous and Bhat (2012) used a spatial panel ordered-response model to analyze the urban land-use development intensity patterns; Ferdous et al. (2011) applied CML in a hazard-based specification to study the nonmotorized mode use; and Sener and Bhat (2011) applied CML to a copula-based spatial unordered response model structure to analyze teenagers' activity participation rates.

81.4.3 Bayesian Approach

Another technique used popularly to estimate complex spatial models is the Bayesian approach. In contrast to classical statistical analysis, the Bayesian approach is rather straightforward in both model estimation and result interpretation. Bayesian framework relies on a set of conditional distributions to deduce each parameter's marginal distribution. In this way, it decomposes multilayered probability specifications into a set of simpler subproblems. A Bayesian approach also allows direct interpretation of parameter estimates and probabilities, since it yields estimates of parameter distributions (rather than classical point estimates, which rely on asymptotic normality). Another advantage of Bayesian approach is that one can integrate established intuition and experience with new information found in sample data through the use of prior or constrained distributions. Thanks to its many advantages,

the Bayesian framework has recently attracted the attention of numerous regional scientists (LeSage and Pace 2009; Smith and LeSage 2004; Wang et al. 2012; Wang and Kockelman 2009). Essentially, Bayesian approaches rest on the basic property of conditional probability known as Bayes' rule:

$$\pi(\beta|y, x) = \frac{\pi(\beta|X)\pi(y|\beta, X)}{\pi(y|X)} \propto \pi(\beta|X)\pi(y|\beta, X) \quad (81.25)$$

Normally, the explanatory variables X are irrelevant to the parameters (as in most cases), so $\pi(\beta|X) = \pi(\beta)$. This is known as the prior, or prior distribution of the random parameters β . One can incorporate intuition and/or experience in this prior distribution. $\pi(y|\beta, X)$ is the likelihood function of y given X and β . Apparently, Bayesian methods integrate information from two sources: one's beliefs and sample data. Together, they lead to updated information on β , producing a posterior distribution of β , which is denoted as $\pi(\beta|y, X)$.

For example, in a SEM multinomial discrete response setting, where Eq. (81.7) is further allowed to incorporate spatial effects in error terms so that

$$\varepsilon_{ik} = \lambda_k \sum_{\substack{j=1 \\ j \neq i}}^N w_{ijk} \varepsilon_{jk} + \theta_{ik} \quad (81.26)$$

where λ_k indicates the spatial coefficient. Let the error terms θ_{ik} between alternatives follow multivariate normal distribution $\theta_i \sim N(0, B)$; then the covariance matrix of ε is Ω , a function of λ_k and B . According to Eq. (81.24), the joint posterior distribution for all parameters can be expressed as

$$p(y^\rightarrow, \beta, \lambda_k, B|y, X) \propto p(y|y^\rightarrow)p(y^\rightarrow|\beta, \lambda_k, B)\pi(\beta)\pi(\lambda_k)\pi(B) \quad (81.27)$$

As the equations only require proportionality, the conditional posterior distributions of all parameters can be derived by extracting only items that contain them, as follows:

$$p(\beta|\dots) \propto p(y^\rightarrow|\beta, \lambda_k, B)\pi(\beta) \quad (81.28)$$

$$p(y^\rightarrow|\dots) \propto p(y|y^\rightarrow)p(y^\rightarrow|\beta, \lambda_k, B) \quad (81.29)$$

$$p(B|\dots) \propto P(y^\rightarrow|\beta, \lambda_k, B)\pi(B) \quad (81.30)$$

$$p(\lambda_k|\dots) \propto p(y^\rightarrow|\beta, \lambda_k, B)\pi(\lambda_k) \quad (81.31)$$

The $p(y|y^\rightarrow)$ and $p(y^\rightarrow|\beta, \lambda_k, B)$ can be derived easily from the model specification. The challenge is to choose appropriate priors. When no established intuition or experience is available, diffuse (also called noninformative or flat) priors are often preferred, reflecting the notion of “letting the data speak for themselves.”

For example, if it is assumed that β has normal priors with hyperparameters β_O and Θ_β , that is,

$$\pi(\beta) \sim N(\beta_O, \Theta_\beta) \quad (81.32)$$

Then the conditional posterior distribution can be derived as

$$p(\beta | \dots) \propto \exp\left(-\frac{1}{2}([y^\rightarrow - X\beta]'\Omega^{-1}[y^\rightarrow - X\beta] + [\beta - \beta_o]'\Theta_\beta^{-1}[\beta - \beta_o])\right) \quad (81.33)$$

It can be observed that such a posterior distribution is essentially a weighted average of β 's prior distribution and sample data information. The weights are the *inverse* of the variance-covariance matrices or associated “uncertainty” levels. Conditional posterior distributions of other parameters can be derived in similar ways. Once derived, they can be estimated using Monte Carlo Markov Chain (MCMC) simulation technique, where the values of parameters are drawn sequentially from these distributions, leading to estimation of each parameter's distribution (see Mills and Parent, ► [Chap. 79](#), “Bayesian MCMC Estimation”).

81.5 Conclusions

Variables can be limited, censored, or truncated for various reasons. Although the forms of models for handling these variables are different, they share one thing in common: the use of latent dependent variables. The observed dependent variables are often expressed as nonlinear functions of the latent variables, which are usually linear functions of explanatory variables. By using latent variables as the bridge, it is convenient to extend the spatial considerations, commonly seen in linear models, into these nonlinear model settings. When the spatial variation of the data generation process is of major concern, geographically weighted regression is preferred. When interest is mainly on prediction or analysis of primary variables, spatial filtering is a powerful tool. When spatial effects need to be explicitly identified and interpreted, spatial regression models are a good choice. Of course, these specifications can also be combined when both spatial variation and dependency need to be addressed.

When the limited and censored dependent variable models are further complicated by the consideration of spatial effects, since the model likelihood becomes multilayered and intractable. In order to obtain consistent and efficient estimators, the selection of estimation techniques is a key consideration. The popular techniques used in practice include maximum simulated likelihood estimation, composite marginal likelihood estimation, and Bayesian approaches. Each of these approaches has its advantages and limitations. The selection of estimation approaches needs to take into account the underlying spatial effects, sample size, model form, and available information. For example, when spatial autocorrelation decays fast, composite marginal likelihood performs well. When established intuition needs to be integrated, the Bayesian approach is proved to be a good framework.

References

- Anselin L (2003) Spatial externalities, spatial multipliers, and spatial econometrics. *Int Reg Sci Rev* 26(2):153–166
- Atkinson PM, German SE, Sear DA, Clark MJ (2003) Exploring the relations between riverbank erosion and geomorphological controls using geographically weighted logistic regression. *Geogr Anal* 35(1):58–82
- Beron KJ, Vijverberg WPM (1999) Probit in a spatial context: a Monte Carlo analysis. In: Anselin L, Florax R, Rey S (eds) *Advances in spatial econometrics, methodology, tools and applications*. Springer, Berlin/Heidelberg/New York, pp 169–196
- Bhat CR (2011) The maximum approximated composite marginal likelihood (MACML) estimation of multinomial probit-based unordered response choice models. *Transp Res: Part B* 45(7):923–939
- Dray S, Legendre P, Peres-Neto PR (2006) Spatial modelling: a comprehensive framework for principal coordinate analysis of neighbour matrices (PCNM). *Ecol Model* 196(3–4):483–493
- Dugundji E, Walker J (2005) Discrete choice with social and spatial network interdependencies: an empirical example using mixed generalized extreme value models with field and panel effects. *Transp Res Rec: J Transp Res Board* 1921(1):70–78
- Ferdous N, Bhat CR (2012) Spatial panel ordered-response model with application to the analysis of urban land use development intensity patterns. Working paper. The University of Texas at Austin. <http://amonline.trb.org/lsmqfv/lsmqfv/1>. Accessed 1 Mar 2012
- Ferdous N, Pendyala R, Bhat C, Konduri K (2011) Modeling the influence of family, social context, and spatial proximity on use of nonmotorized transport mode. *Transp Res Rec: J Transp Res Board* 2230(1):111–120
- Fotheringham S (2003) Geographically weighted regression: the analysis of spatially varying relationships. Wiley, West Sussex
- Getis A (1995) Spatial filtering in a regression framework: experiments on regional inequality, government expenditures, and urban crime. In: *New directions in spatial econometrics*. Springer, Berlin/Heidelberg/New York, pp 172–188
- Greene WH (2002) *Econometric analysis*, 5th edn. Prentice Hall, Upper Saddle River
- Griffith DA (2000) Eigenfunction properties and approximations of selected incidence matrices employed in spatial analyses. *Linear Algebra Appl* 321(1–3):95–112
- Klier T, McMillen DP (2008) Clustering of auto supplier plants in the U.S.: GMM spatial logit for large samples. *J Bus Econ Stat* 26(4):460–471
- LeSage JP (1999) Applied econometrics using MATLAB, <http://www.spatial-econometrics.com/html/mbook.pdf>. Accessed 1 Mar 2012
- LeSage JP, Pace RK (2009) Introduction to spatial econometrics. CRC Press/Taylor & Francis Group, Boca Raton
- Luo J, Wei YHD (2009) Modeling spatial variations of urban growth patterns in Chinese cities: the case of Nanjing. *Landscape Urban Plan* 91(2):51–64
- McFadden D (1980) Econometric models for probabilistic choice among products. *J Bus* 53(3): S13–S29
- McMillen DP (1995) Spatial effects in probit models: a Monte Carlo investigation. In: Anselin L, Florax R (eds) *New directions in spatial econometrics*. Springer, Berlin/Heidelberg/New York, pp 189–228
- McMillen DP, McDonald JF (1999) Land use before zoning: the case of 1920's Chicago. *Reg Sci Urban Econ* 29(4):473–489
- Pace RK, LeSage JP (2011) Fast simulated maximum likelihood estimation of the spatial probit model capable of handling large samples. <http://ssrn.com/abstract=1966039>. Accessed 15 Feb 2012
- Paletti R, Bhat CR (2011) The composite marginal likelihood (CML) estimation of panel ordered-response models. Working paper. The University of Texas at Austin. http://www.caee.utexas.edu/prof/bhat/ABSTRACTS/CML_Paper_27July2010.pdf. Accessed 1 Jan 2012

- Pinkse J, Slade ME (1998) Contracting in space: an application of spatial statistics to discrete-choice models. *J Econometrics* 85(1):125–154
- Sener I, Bhat C (2011) Flexible spatial dependence structures for unordered multinomial choice models: formulation and application to teenagers' activity participation. *Transportation* 39(13):657–683
- Smith TE, LeSage JP (2004) A Bayesian probit model with spatial dependencies. In: Pace RK, LeSage JP (eds) *Advances in econometrics: spatial and spatiotemporal econometric*, vol 18. Elsevier, Oxford, pp 127–160
- Train K (2003) Discrete choice methods with simulation. Cambridge University Press, New York
- Varin C (2008) On composite marginal likelihoods. *AStA Adv Stat Anal* 92(1):1–28
- Vijverberg WPM (1997) Monte Carlo evaluation of multivariate normal probabilities. *J Econometrics* 76(1–2):281–307
- Wang X, Kockelman K (2008) Maximum simulated likelihood estimation with correlated observations: a comparison of simulation techniques. In: Sloboda B (ed) *Transportation statistics*. J. D Ross Publishing, Fort Lauderdale, pp 173–194
- Wang X, Kockelman K (2009) Bayesian inference for ordered response data with a dynamic spatial-ordered probit model. *J Reg Sci* 49(5):877–913
- Wang X, Kockelman K, Lemp J (2012) The dynamic spatial multinomial probit model: analysis of land use change using parcel-level data. *J Transp Geogr* 24:77–88

J. Paul Elhorst

Contents

82.1	Introduction	1637
82.2	Linear Spatial Dependence Models for Cross-Sectional Data	1638
82.3	Linear Spatial Dependence Models for Panel Data	1641
82.4	Dynamic Linear Spatial Dependence Models for Panel Data	1643
82.5	Direct and Spatial Spillover Effects	1645
82.6	Empirical Illustration	1646
82.7	Conclusion	1650
	References	1651

Abstract

This chapter provides a survey of the existing literature on spatial panel data models. Both static and dynamic models will be considered. The chapter also demonstrates that spatial econometric models that include lags of the dependent variable and of the independent variables in both space and time provide a useful tool to quantify the magnitude of direct and indirect effects, both in the short term and long term. Direct effects can be used to test the hypothesis as to whether a particular variable has a significant effect on the dependent variable in its own economy and indirect effects to test the hypothesis whether spatial spillovers exist. To illustrate these models and their effects estimates, a demand model for cigarettes is estimated based on panel data from 46 US states over the period 1963–1992.

J.P. Elhorst

Department of Economics, Econometrics and Finance, University of Groningen, Groningen,
The Netherlands
e-mail: j.p.elhorst@rug.nl

82.1 Introduction

Spatial econometrics deals with interaction effects among geographical units, such as zip codes, neighborhoods, municipalities, counties, regions, states, or countries. Examples are economic growth rates of OECD countries over T years, monthly unemployment rates of EU regions in the last decade, and annual tax rate changes of all jurisdictions in a country since the last election. Spatial econometric models can also be used to explain the behavior of economic agents other than geographical units, such as individuals, firms, or governments, but this type of research is still in its infancy. Examples are research productivity of N universities located in a particular country and consumption of the representative consumer in each zip code of the trade area of a commercial firm.

In modeling terms, three different types of interaction effects can be distinguished: endogenous interaction effects among the dependent variable (Y), exogenous interaction effects among the independent variables (X), and interaction effects among the error terms (ϵ). Originally, the central focus of spatial econometrics has been on one type of interaction effect in a single-equation cross-sectional setting (► Chap. 66, “Exploratory Spatial Data Analysis”). Usually, the point estimate of the coefficient of this interaction effect was used to test the hypothesis as to whether spatial spillover effects exist. Most of the work was inspired by research questions arising in regional science and economic geography, where the units of observations are geographically determined and the structure of the dependence among these units can somehow be related to location and distance. However, more recently, the focus has shifted to models with more than one type of interaction effects, to panel data, and to the marginal effects of the explanatory variables in the model rather than the point estimates of the interaction effects.

In this chapter, we review and organize this recent literature. In Sect. 82.2, we present the linear regression model with spatial interaction effects for cross-sectional data and in Sect. 82.3 its extension to panel data. In Sect. 82.4, the latter model is further extended to include dynamic effects in both space and time. In Sect. 82.5, we provide so-called effects estimates (► Chap. 73, “Geographically Weighted Regression”), which are required for making correct inferences regarding the effect of independent variables on the dependent variable. In Sect. 82.6, we estimate a demand model for cigarettes based on panel data from 46 US states over the period 1963–1992 to empirically illustrate the different models. This data set is taken from Baltagi (2005) and has been used for illustration purposes in other studies too. Finally, we conclude this chapter with a number of important implications for econometric modeling of relationships based on spatial panel data.

82.2 Linear Spatial Dependence Models for Cross-Sectional Data

The standard approach in most empirical work is to start with a nonspatial linear regression model and then to test whether or not the model needs to be extended

with spatial interaction effects. This approach is known as the specific-to-general approach. The nonspatial linear regression model takes the form

$$Y = \alpha t_N + X\beta + \varepsilon \quad (82.1)$$

where Y denotes an $N \times 1$ vector consisting of one observation on the dependent variable for every unit in the sample ($i = 1, \dots, N$), t_N is an $N \times 1$ vector of ones associated with the constant term parameter α , X denotes an $N \times K$ matrix of exogenous explanatory variables with parameters β contained in a $K \times 1$ vector, and $\varepsilon = (\varepsilon_1, \dots, \varepsilon_N)^T$ is a vector of disturbance terms. The ε_i are independently and identically distributed error terms for all i with zero mean and variance σ^2 . Since the linear regression model is commonly estimated by ordinary least squares (OLS), it is often labeled the OLS model. Furthermore, even though the OLS model in most studies focusing on spatial interaction effects is rejected in favor of a more general model, its results often serve as a benchmark.

The opposite approach is to start with a more general model that nests a series of simpler models, which would ideally represent all alternative economic hypotheses requiring consideration. Generally, three different types of interaction effects may explain why an observation associated with a specific location may be dependent on observations at other locations:

- (i) Endogenous interaction effects, where the decision of a particular unit A (or its economic decision makers) to behave in some way depends on the decision taken by other units, among which, say, unit B,

$$\text{Dependent variable } y \text{ of unit A} \leftrightarrow \text{Dependent variable } y \text{ of unit B} \quad (82.2)$$

Endogenous interaction effects are typically considered as the formal specification for the equilibrium outcome of a spatial or social interaction process, in which the value of the dependent variable for one agent is jointly determined with that of the neighboring agents. In the empirical literature on strategic interaction among local governments, for example, endogenous interaction effects are theoretically consistent with the situation where taxation and expenditures on public services interact with taxation and expenditures on public services in nearby jurisdictions (Brueckner 2003).

- (ii) Exogenous interaction effects, where the decision of a particular unit to behave in some way depends on independent explanatory variables of the decision taken by other units

$$\text{Independent variable } x \text{ of unit B} \rightarrow \text{Dependent variable } y \text{ of unit A} \quad (82.3)$$

Consider, for example, the savings rate. According to standard economic theory, saving and investment are always equal. People cannot save without investing their money somewhere, and they cannot invest without using somebody's savings. This is true for the world as a whole, but it is not true

for individual economies. Capital can flow across borders; hence the amount an individual economy saves does not have to be the same as the amount it invests. In other words, per capita income in one economy also depends on the savings rates of neighboring economies. It should be stressed that, if the number of independent explanatory variables in a linear regression model is K , the number of exogenous interaction effects might also be K , provided that the intercept is considered as a separate variable. In other words, not only the savings rate but also other explanatory variables may affect per capita income in neighboring economies. It is for this reason that economic growth in both the theoretical and the empirical literature on economic growth and convergence among countries or regions is not only taken to depend on the initial income level and the rates of saving, population growth, technological change, and depreciation in the own economy but also on those of neighboring economies (Ertur and Koch 2007; Elhorst et al. 2010).

- (iii) Interaction effects among the error terms

$$\text{Error term } u \text{ of unit A} \leftrightarrow \text{Error term } u \text{ of unit B} \quad (82.4)$$

Interaction effects among the error terms do not require a theoretical model for a spatial or social interaction process but, instead, are consistent with a situation where determinants of the dependent variable omitted from the model are spatially autocorrelated and with a situation where unobserved shocks follow a spatial pattern. Interaction effects among the error terms may also be interpreted to reflect a mechanism to correct rent-seeking politicians for unanticipated fiscal policy changes (Allers and Elhorst 2005).

A full model with all types of interaction effects takes the form

$$Y = \rho WY + \alpha \iota_N + X\beta + WX\theta + u \quad (82.5a)$$

$$u = \lambda Wu + \epsilon \quad (82.5b)$$

where the variable WY denotes the endogenous interaction effects among the dependent variables, WX the exogenous interaction effects among the independent variables, and Wu the interaction effects among the disturbance terms of the different units. ρ is called the spatial autoregressive coefficient and λ the spatial autocorrelation coefficient, while θ , just as β , represents a $K \times 1$ vector of fixed but unknown parameters. W is a nonnegative $N \times N$ matrix of known constants describing the arrangement of the units in the sample. Its diagonal elements are set to zero by assumption, since no unit can be viewed as its own neighbor.

Three methods have been developed to estimate models that include interaction effects. One is based on maximum likelihood (ML) or quasi-maximum likelihood (QML), one on instrumental variables or generalized method of moments (IV/GMM), and one on the Bayesian Markov Chain Monte Carlo (MCMC)

approach. QML and IV/GMM estimators are different in that they do not rely on the assumption of normality of the disturbances. Detailed descriptions of these estimation techniques can be found in Anselin (1988), Lee (2004), Kelejian and Prucha (1998), LeSage and Pace (2009), ► Chap. 67, “Spatial Clustering and Autocorrelation in Health Events,” ► Chap. 69, “Spatial Dynamics and Space-Time Data Analysis,” and ► Chap. 72, “Bayesian Spatial Statistical Modeling”.

Technically, there are no obstacles to estimating a model with interaction effects among the dependent variable, the independent variables, and the disturbance terms. Often, however, the parameters cannot be interpreted in a meaningful way since the different types of interaction effects cannot be distinguished from each other. Lee et al. (2010) prove that there is at least one spatial weights matrix so that all parameters are identified. They consider G groups, each consisting of N_g cross-sectional units, and assume that the elements of the spatial weights matrix measuring the interaction effects are $w_{ij} = 1/(N_g - 1)$ if units i and j belong to the same group (except if $i = j$) and zero otherwise. Starting with this spatial weights matrix, it is shown that the parameters are identified either if both N and N_g tend to infinity, with at least two units in each group, or if the number of units in each group does not tend to infinity faster than or equal to the number of groups. Whether the parameters are also identified for other specifications of the spatial weights matrix still needs to be investigated. If not, the best option is to exclude the spatially autocorrelated error term and to consider a model with endogenous and exogenous interaction effects “only.” This model is known as the spatial Durbin model. Only is put in quotation marks, because this model covers $K + 1$ of the $K + 2$ potential interaction effects. According to LeSage and Pace (2009, pp. 155–158), the cost of ignoring spatial dependence in the dependent variable and/or in the independent variables is relatively high since the econometric literature has pointed out that if one or more relevant explanatory variables are omitted from a regression equation, the estimator of the coefficients for the remaining variables is biased and inconsistent. In contrast, ignoring spatial dependence in the disturbances, if present, will only cause a loss of efficiency.

82.3 Linear Spatial Dependence Models for Panel Data

In recent years, the spatial econometric literature has exhibited a growing interest in the specification and estimation of econometric relationships based on spatial panels. This interest can be partly explained by the increased availability of more data sets in which a number of spatial units are followed over time and partly by the fact that panel data offer researchers extended modeling possibilities as compared to the single-equation cross-sectional setting. Panel data are generally more informative, and they contain more variation and less collinearity among the variables. The use of panel data results in a greater availability of degrees of freedom and hence increases efficiency in the estimation. Panel data also allow for the specification of more complicated behavioral hypotheses, including effects that cannot be addressed using pure cross-sectional data (see Baltagi 2005 and the references therein).

The extension of the spatial econometric model, presented in Eqs. (82.5a, 82.5b), for a cross section of N observations to a space-time model for a panel of N observations over T time periods is obtained by adding a subscript t , which runs from 1 to T , to the variables and the error terms of that model

$$Y_t = \rho W Y_t + \alpha_{tN} + X_t \beta + W X_t \theta + u_t \quad (82.6a)$$

$$u_t = \lambda W u_t + \varepsilon_t \quad (82.6b)$$

This model can be estimated along the same lines as the cross-sectional model, provided that all notations are adjusted from one cross section to T cross sections of N observations.

However, the main objection to pooling the data like this is that the resulting model does not account for spatial and temporal heterogeneity. Spatial units are likely to differ in their background variables, which are usually space-specific time-invariant variables that do affect the dependent variable, but which are difficult to measure or hard to obtain. Examples of such variables abound: one spatial unit is located at the seaside, the other just at the border; one spatial unit is a rural area located in the periphery of a country, the other an urban area located in the center; norms and values regarding labor, crime, and religion in one spatial unit might differ substantially from those in another unit. Failing to account for these variables increases the risk of obtaining biased estimation results. One remedy is to introduce a variable intercept μ_i representing the effect of the omitted variables that are peculiar to each spatial unit considered. In sum, spatial-specific effects control for all time-invariant variables whose omission could bias the estimates in a typical cross-sectional study. Similarly, the justification for adding time period-specific effects ξ_t is that they control for all spatial-invariant variables whose omission could bias the estimates in a typical time-series study (Baltagi 2005). Examples of such variables also exist: one year is marked by economic recession, the other by a boom; changes in legislation or government policy can significantly affect the functioning of an economy as from the date of implementation, as a result of which before and after observations might be significantly different from one another.

The space-time model in Eqs. (82.6a, 82.6b) extended with spatial-specific and time-period-specific effects reads as

$$Y_t = \rho W Y_t + \alpha_{tN} + X_t \beta + W X_t \theta + \mu + \xi_{tN} + u_t \quad (82.7a)$$

$$u_t = \lambda W u_t + \varepsilon_t \quad (82.7b)$$

where $\mu = (\mu_1, \dots, \mu_N)^T$. The spatial- and time period-specific effects may be treated as fixed effects or as random effects. In the fixed effects model, a dummy variable is introduced for each spatial unit and for each time period (except one to avoid perfect multicollinearity), while in the random effects model, μ_i and ξ_t are

treated as random variables that are independently and identically distributed with zero mean and variance σ_{μ}^2 and σ_{ξ}^2 , respectively. Furthermore, it is assumed that the random variables μ_i , ξ_t , and ε_{it} are independent of each other.

The estimation of static spatial panel data models is extensively discussed in Elhorst (2003, 2010a) and Lee and Yu (2010a). The first presents the ML estimator of the spatial lag model and of the spatial error model extended to include fixed effect or random effects. Further note that the spatial Durbin model can be estimated as a spatial lag model with explanatory variables $[X \; WX]$ instead of X . The response parameters of the fixed effects models can be estimated by concentrating out the fixed effects first, called demeaning (see Baltagi (2005) for mathematical details). The resulting equation can then be estimated by the ML estimation procedure developed by Anselin (1988) for the spatial lag model, provided that this procedure is generalized from one single cross section of N observations to T cross sections of N observations. The estimation of the random effects model is somewhat more complicated.

Lee and Yu (2010a) show that the ML estimator of the spatial lag and of the spatial error model with spatial fixed effects, as set out in Elhorst (2003, 2010a), will yield an inconsistent parameter estimate of the variance parameter (σ^2) if N is large and T is small and inconsistent estimates of all parameters of the spatial lag and of the spatial error model with spatial and time-period fixed effects if both N and T are large. To correct for this, they propose a simple bias correction procedure based on the parameter estimates of the uncorrected approach. Matlab routines for both the fixed effects and the random effects spatial lag model, as well as the fixed effects and the random effects spatial error model are provided at www.regroningen.nl/elhorst. Recently, these routines have been updated for the bias correction procedure of Lee and Yu (2010a).

82.4 Dynamic Linear Spatial Dependence Models for Panel Data

To make the spatial panel data model, presented in Eqs. (82.7a, 82.7b), dynamic, one might add time lags of the variables Y_t and WY_t to get

$$Y_t = \tau Y_{t-1} + \rho WY_t + \eta WY_{t-1} + \alpha_{1N} + X_t \beta + WX_t \theta + \mu + \xi_t \iota_N + u_t \quad (82.8)$$

This model is known as a dynamic spatial Durbin model (Debarsy et al. 2012). Similarly, one might consider time lags of the variables X_t and WX_t and of the error terms u_t and Wu_t , but according to Anselin et al. (2008), the parameters of such a model will not be identified.

Three methods have been developed in the literature to estimate models that have mixed dynamics in both space and time. One method is to bias-correct the maximum likelihood (ML) or quasi-maximum likelihood (QML) estimator, one method is based on instrumental variables or generalized method of moments (IV/GMM), and one method utilizes the Bayesian Markov Chain Monte Carlo (MCMC) approach.

Yu et al. (2008) construct a bias-corrected estimator for a dynamic model (Y_{t-1} , WY_t and WY_{t-1}) with spatial fixed effects. Lee and Yu (2010b) extend this study to include time-period fixed effects. They first estimate the model by the ML estimator for the spatial lag model with spatial (and time-period) fixed effects, conditional upon the first observation of every spatial unit in the sample due to the regressors Y_{t-1} and WY_{t-1} . Next, they provide a rigorous asymptotic theory for their ML estimator and suggest a bias-corrected ML estimator when both the number of spatial units (N) and the number of time points (T) in the sample go to infinity such that the limit between N and T exists and is bounded between zero and infinity ($0 < \lim(N/T) < \infty$). In the words of Lee and Yu (2010c, p. 2), this condition implies that “ $T \rightarrow \infty$ where T cannot be too small relative to N .” The bias correction is derived for both normally distributed error terms (ML) and for error terms that do not rely on the normality assumption (QML). In the latter case, the first four moments are required. Finally, it is to be noted that this bias-corrected ML estimator can also be used when either the variable Y_{t-1} or the variable WY_{t-1} is eliminated from the model.

Elhorst (2010b) investigates the small sample properties of the bias-corrected ML estimator. For this purpose, he extends the unconditional ML estimator proposed by Hsiao et al. (2002) with the variable WY_t , as well as the Bhargava and Sargan (1983) approximation that is used to determine the expected value and the variance of the first first-differenced observations in the sample. One of his conclusions is that the parameter estimate ρ of the variable WY_t is still considerably biased when using this unconditional ML estimator. However, if the parameter estimate ρ is based on the bias-corrected ML estimator and the other parameters, given ρ , on the unconditional ML estimator, then this so-called mixed estimator outperforms the bias-corrected estimator of Yu et al. (2008) for small values of T ($T = 5$).

A couple of studies have considered IV/GMM estimators, building on previous work of Arellano and Bond (1991) and Blundell and Bond (1998). Elhorst (2010b) extends the Arellano and Bond difference GMM estimator to include endogenous interaction effects and finds that this estimator can still be severely biased, especially with respect to the parameter estimate ρ of the variable WY_t . He notes a bias of 0.061. The explanation for this can be found in Lee and Yu (2010c). They find that a 2SLS estimator like the Arellano and Bond GMM estimator which is based on lagged values of Y_{t-1} , WY_{t-1} , X_t , and WX_t is not consistent due to too many moments and that the dominant bias is caused by the endogeneity of the variable WY_t rather than the variable Y_{t-1} . To avoid these problems, they propose an optimal GMM estimator based on linear moment conditions, which are standard, and quadratic moment conditions, which are implied by the variable WY_t , and therefore not standard in dynamic panel data models. They prove that this GMM estimator is consistent, also when T is small relative to N .

Parent and LeSage (2010, 2011) point out that the Bayesian MCMC approach considers conditional distributions of each parameter of interest conditional on the others, which leads to some computational simplification. Just as Elhorst (2010b), they treat the first-period cross section as endogenous, using the Bhargava and

Sargan (1983) approximation, and find that the correct treatment of the initial observations (endogenous instead of exogenous) is important, especially in cases when T is small.

82.5 Direct and Spatial Spillover Effects

Many empirical studies use point estimates of one or more spatial regression model specifications to test the hypothesis as to whether or not spatial spillovers exist. One of the key contributions of LeSage and Pace's book (2009, p. 74) is the observation that this may lead to erroneous conclusions and that a partial derivative interpretation of the impact from changes to the variables of different model specifications represents a more valid basis for testing this hypothesis.

By rewriting the spatial econometric model with dynamic effects in space and time in Eq. (82.8) as

$$Y_t = (I - \rho W)^{-1}(\tau I + \eta W)Y_{t-1} + (I - \rho W)^{-1}(X_t\beta + WX_t\theta) + R \quad (82.9)$$

where R is a rest term containing the intercept and the error terms, the matrix of partial derivatives of the expected value of Y with respect to the k th explanatory variable of X in unit 1 up to unit N at a particular point in time can be seen to be

$$\begin{bmatrix} \frac{\partial E(Y)}{\partial x_{1k}} & \dots & \frac{\partial E(Y)}{\partial x_{Nk}} \end{bmatrix}_t = (I - \rho W)^{-1}[\beta_k I_N + \theta_k W] \quad (82.10)$$

These partial derivatives denote the effect of a change of a particular explanatory variable in a particular spatial unit on the dependent variable of all other units in the *short term*. Similarly, the *long-term* effects can be seen to be

$$\begin{bmatrix} \frac{\partial E(Y)}{\partial x_{1k}} & \dots & \frac{\partial E(Y)}{\partial x_{Nk}} \end{bmatrix} = [(1 - \tau)I - (\rho + \eta)W]^{-1}[\beta_k I_N + \theta_k W] \quad (82.11)$$

LeSage and Pace (2009) and Debarsy et al. (2012) define the direct effect as the average of the diagonal elements of the matrix on the right-hand side of Eq. (82.10) or Eq. (82.11) and the indirect effect as the average of either the row sums or the column sums of the non-diagonal elements of these matrices (since the numerical magnitudes of these two calculations of the indirect effect are the same, it does not matter which one is used). The outcomes are independent from the time index; this explains why the right-hand sides of these equations do not contain the symbol t . The expressions in Eqs. (82.10) and (82.11) also show that short-term indirect effects do not occur if both $\rho = 0$ and $\theta_k = 0$, while long-term indirect effects do not occur if both $\rho = -\eta$ and $\theta_k = 0$.

Using the expressions in Eqs. (82.10) and (82.11), it is also possible to indicate the disadvantages of certain parameter restrictions put forward in previous studies (see Elhorst (2012) for an overview). The disadvantage of not considering

exogenous interaction effects (WX_t), that is, by imposing the restriction $\theta = 0$, is that the ratio between the indirect effect and the direct effect becomes the same for every explanatory variable; if this ratio happens to be p percent for one variable, it is also p percent for any other variable. The disadvantage of excluding contemporaneous endogenous interaction effects (WY_t) by imposing the restriction $\rho = 0$ is that the matrix $(I - \rho W)^{-1}$ degenerates to the identity matrix, as a result of which the indirect effects in the short term only depend on θ . This loss of flexibility makes the model less suitable for empirical research focusing on short-term effects. The disadvantage of imposing the restriction $\eta = -\tau\rho$ is that the ratio between the indirect effect and the direct effect of a particular explanatory variable remains constant over time; if this ratio happens to be p percent for one variable in the short term, it is also p percent in the long term. Note that this restriction implies that η , the parameter associated with lagged endogenous interaction effects, is equal to $-\tau\rho$, the two parameters of the dependent variables respectively lagged in time and lagged in space. In other words, just as the two previous restrictions, it eliminates one type of interaction effects. Finally, if lagged endogenous interaction effects (WY_{t-1}) are eliminated from the model, that is, $\eta = 0$, no prior restrictions are imposed on the direct and indirect effects estimates, even though it is clear that still some flexibility of the model will be lost.

82.6 Empirical Illustration

Baltagi and Li (2004) estimate a demand model for cigarettes based on a panel from 46 US states in which real per capita sales of cigarettes by persons of smoking age (14 years and older) measured in packs of cigarettes per capita (C_{it}) is regressed on the average retail price of a pack of cigarettes measured in real terms (P_{it}) and on real per capita disposable income (Y_{it}). Moreover, all variables are taken in logs. Whereas Baltagi and Li (2004) use the first 25 years for estimation to reserve data for out-of-sample forecasts, we use the full data set covering the period 1963–1992. This data set can be downloaded freely from www.wiley.co.uk/baltagi/, while an adapted version is available at www.regroningen.nl/elhorst. More details, as well as reasons to include state-specific effects (μ_i) and time-specific effects (ξ_t), are given in Baltagi (2005). The spatial weights matrix is specified as a row-normalized binary contiguity matrix whose elements are one if two states share a common border and zero otherwise.

Column (1) of Table 82.1 reports the estimation results when adopting a non-dynamic spatial Durbin model without spatial and time-period fixed effects and column (2) when including these effects. To investigate whether or not the fixed effects are jointly significant, one may test the hypothesis $H_0: \mu_1 = \dots = \mu_N = \xi_1 = \dots = \xi_T = \alpha$, where α is the intercept of the model without fixed effects. To test this (null) hypothesis, one may perform a likelihood-ratio (LR) test, which is based on the log-likelihood function values of both models. The number of degrees of freedom is equal to the number of restrictions that needs to be imposed on the fixed effects to get one overall intercept, which in this particular case is $N + T - 1$. The outcome of this test $(2 \times (1691.4 - 475.5)) = 2431.8$ with

Table 82.1 Estimation results of cigarette demand using different model specifications

Determinants	(1)	(2)	(3)
	Non-dynamic spatial Durbin model no fixed effects	Non-dynamic spatial Durbin model with fixed effects	Dynamic spatial Durbin model with fixed effects
Intercept	2.631 (15.82)		
Log(C) ₋₁		0.865 (65.04)	
W*Log(C)	0.337 (11.09)	0.264 (8.25)	0.076 (2.00)
W*Log(C) ₋₁			-0.015 (-0.29)
Log(P)	-1.251 (-21.80)	-1.001 (-24.36)	-0.266 (-13.19)
Log(Y)	0.554 (14.96)	0.603 (10.27)	0.100 (4.16)
W*Log(P)	0.780 (11.15)	0.093 (1.13)	0.170 (3.66)
W*Log(Y)	-0.444 (11.09)	-0.314 (-3.93)	-0.022 (-0.87)
R ²	0.435	0.902	0.977
LogL	475.5	1691.4	2623.3

Notes: t-values in parentheses

$N + T - 1 = 46 + 30 - 1 = 75$ df) justifies the extension of the model with spatial and time-period effects. Note that one may also separately test for the inclusion of spatial fixed effects and time-period fixed effects.

It is to be noted that the coefficient of any variable that does not change over time or only a little cannot be estimated when controlling for spatial fixed effects. Similarly, the coefficient of any variable that does not change across space or only a little cannot be estimated when controlling for time-period fixed effects. For many empirical studies, this is a reason not to control for fixed effects, for example, because such time-invariant or space-invariant variables are the main focus of the analysis. However, if one or more relevant explanatory variables are omitted from the regression equation, when they should be included, the estimator of the coefficients of the remaining variables is biased and inconsistent. This also holds true for fixed effects and is known as the omitted regressor bias.

Instead of fixed effects, we can also treat μ and ξ as random effects. Hausman specification test can then be used to test the random effects model against the fixed effects model. However, whether the random effects model is an appropriate specification if the population may be said to be sampled exhaustively, such as all counties of a state or all regions in a country, remains controversial. A detailed discussion of this issue can be found in Elhorst (2012).

The main shortcoming of a non-dynamic spatial Durbin model is that it cannot be used to calculate short-term effects estimates of the explanatory variables. This is made clear in Table 82.2, which reports the corresponding effects estimates of the models presented in Table 82.1; since a non-dynamic model only produces long-term effects estimates, the cells reporting short-term effects estimates are left empty.

The direct effects estimates of the two explanatory variables reported in column (2) of Table 82.2 are significantly different from zero and have the expected signs. Higher prices restrain people from smoking, while higher-income levels have

Table 82.2 Effects estimates of cigarette demand using different model specifications

Determinants	(1)	(2)	(3)
	Non-dynamic spatial Durbin model no fixed effects	Non-dynamic spatial Durbin model with fixed effects	Dynamic spatial Durbin model with fixed effects
Short-term direct effect Log(P)			-0.262 (-11.48)
Short-term indirect effect Log(P)			0.160 (3.49)
Short-term direct effect Log(Y)			0.099 (3.36)
Short-term indirect effect Log(Y)			-0.018 (-0.45)
Long-term direct effect Log(P)	-1.216 (-23.39)	-1.013 (-24.73)	-1.931 (-9.59)
Long-term indirect effect Log(P)	0.508 (7.27)	-0.220 (-2.26)	0.610 (0.98)
Long-term direct effect Log(Y)	0.530 (15.48)	0.594 (10.45)	0.770 (3.55)
Long-term indirect effect Log(Y)	-0.366 (-7.47)	-0.197 (-2.15)	0.345 (0.48)

Notes: t-values in parentheses

a positive effect on cigarette demand. The price elasticity amounts to -1.013 and the income elasticity to 0.594 . Note that these direct effects estimates are different from the coefficient estimates of -1.001 and 0.603 reported in column (2) of [Table 82.1](#) due to feedback effects that arise as a result of impacts passing through neighboring states and back to the states themselves.

The spatial spillover effects (indirect effects estimates) of both variables are negative and significant. Own-state price increases will restrain people not only from buying cigarettes in their own state but to a limited extent also from buying cigarettes in neighboring states (elasticity -0.220). By contrast, whereas an income increase has a positive effect on cigarette consumption in the own state, it has a negative effect in neighboring states. We come back to this result below. Further note that the non-dynamic spatial Durbin model without spatial and time-period effects indicates a positive rather than a negative spatial spillover effect of price increases and that only a positive outcome would be consistent with Baltagi and Levin (1992), who found that price increases in a particular state – due to tax increases meant to reduce cigarette smoking and to limit the exposure of non-smokers to cigarette smoke – encourage consumers in that state to search for cheaper cigarettes in neighboring states. However, there are two reasons why this comparison is invalid. First, whereas Baltagi and Levin's (1992) model is dynamic, it is not spatial. They do consider the price of cigarettes in neighboring states, but not any other spatial interaction effects. Second, whereas our model contains spatial interaction effects, it is not (yet) dynamic. For these reasons, it is interesting to consider the estimation results of our dynamic spatial panel data model.

Column (3) of [Table 82.2](#) reports the direct and indirect effects of the dynamic model, both in the short term and long term. Consistent with microeconomic theory, the short-term direct effects appear to be substantially smaller than the long-term direct effects: -0.262 versus -1.931 for the price variable and 0.099 versus 0.770 for the income variable. This is because it takes time before price and income changes fully settle. The long-term direct effects in the dynamic spatial Durbin

model, on their turn, appear to be greater (in absolute value) than their counterparts in the non-dynamic spatial Durbin model: -1.931 versus -1.013 for the price variable and 0.770 versus 0.594 for the income variable. Apparently, the non-dynamic model underestimates the long-term effects. The short-term spatial spillover effect of a price increase turns out to be positive; the elasticity amounts to 0.160 and is highly significant (t -value, 3.49). This finding is in line with the original finding of Baltagi and Levin (1992) in that a price increase in one state encourages consumers to search for cheaper cigarettes in neighboring states. The negative spatial spillover effect of a price increase we found earlier for the non-dynamic spatial Durbin model demonstrates that a non-dynamic approach falls short here. Although greater and again positive, we do not find empirical evidence that the long-term spatial spillover effect of a price increase is also significant. A similar result is found by Debarsy et al. (2012). It is to be noted that they estimate the parameters of the model by the Bayesian MCMC estimator developed by Parent and LeSage (2010, 2011), whereas we use the bias-corrected ML estimator developed by Lee and Yu (2010b). Furthermore, the spatial weights matrix used in that study is based on lengths of state borders in common between each state and its neighboring states, whereas we use a binary contiguity matrix.

The long-term spatial spillover effect of the income variable derived from the dynamic spatial panel data model appears to be positive, which suggests that an income increase in a particular state has a positive effect on smoking not only in that state itself but also in neighboring states. Furthermore, the spatial spillover effect is smaller than the direct effect, which makes sense since the impact of a change will most likely be larger in the place that instigated the change. However, the spatial spillover effect of an income increase is not significant. A similar result is found by Debarsy et al. (2012). Interestingly, the spatial spillover effect of the income variable in the non-dynamic spatial panel data model appeared to be negative and significant. Apparently, the decision whether to adopt a dynamic or a non-dynamic model represents an important issue. Some researchers prefer simpler models to more complex ones (Occam's razor). One problem of complex models is overfitting, the fact that excessively complex models are affected by statistical noise, whereas simpler models may capture the underlying process better and may thus have better predictive performance. However, if one can trade simplicity for increased explanatory power, the complex model is more likely to be the correct one.

To investigate whether the extension of the non-dynamic model to the dynamic spatial panel data model increases the explanatory power of the model, one may test whether the coefficients of the variables Y_{t-1} and WY_{t-1} are jointly significant using an LR test. The outcome of this test ($2 \times (2,623.3 - 1,691.4) = 1,863.8$ with 2 df) evidently justifies the extension of the model with dynamic effects.

One potential objection to the dynamic spatial Durbin model might be that its parameters are still not identified (Anselin et al. 2008). To investigate this, we carried out a Monte Carlo simulation experiment. The basic idea is to randomly draw (e.g., $1,000$ times) the error terms based on σ^2 of the estimated equation, to generate the dependent variable given this error term and the independent variables

Table 82.3 Identification of parameters

Determinant/parameter	Dynamic spatial Durbin model with fixed effects	
	Original parameter value*	Simulated parameter value
Log(C) ₋₁	0.865	0.864
W*Log(C)	0.076	0.074
W*Log(C) ₋₁	-0.015	-0.005
Log(P)	-0.266	-0.264
Log(Y)	0.100	0.100
W*Log(P)	0.170	0.182
W*Log(Y)	-0.022	-0.025

*Based on column (3) of [Table 82.1](#)

and their coefficient estimates reported in column (3) of [Table 82.1](#), and then to reestimate the model. On average, these results should be similar to those of the “original” parameter estimates. The results reported in [Table 82.3](#) show that the biases in the coefficient estimates, based on 1,000 replications, in this particular case are small. The largest bias is found in the coefficient of the spatial lag of the price variable W*Log(P); its original coefficient is 0.170, while its simulated coefficient is 0.182, which represents a bias of 0.012 or 7.1 % of the original parameter value. Although the impact of this bias on the direct and indirect effects estimates is negligible, a more careful elaboration of the identification issue is the topic of further research.

82.7 Conclusion

Spatial econometric models that include lags of the dependent variable and of the independent variables in both space and time provide a useful tool to quantify the magnitude of direct and indirect effects, both in the short term and long term. A demand model for cigarettes based on panel data from 46 US states over the period 1963–1992 is used to empirically illustrate this. Direct effects should be used to test the hypothesis as to whether a particular variable has a significant effect on the dependent variable in its own economy rather than the coefficient estimate of that variable. Similarly, indirect effects should be used to test whether or not spatial spillovers exist rather than the coefficient estimate of the spatially lagged dependent variable and/or the coefficient estimates of the spatially lagged independent variables.

One difficulty is that it cannot be seen from the coefficient estimates and the corresponding standard errors or t-values (derived from the variance-covariance matrix) whether the direct and indirect effects in a spatial econometric model are significant. This is because these effects are composed of different coefficient estimates according to complex mathematical formulas and the dispersion of these effects depends on the dispersion of all coefficient estimates involved. Fortunately, some individual researchers have made software available at their Web sites programmed in Matlab or R that calculates these effects and

their corresponding standard errors or t-values. Nevertheless, the availability of easier accessible packages such as Stata would probably encourage much more applied researchers to use these kinds of models and to report direct and indirect effects estimates in addition to the point estimates of the parameters of the model. This is important since eventually only these effects estimates should be used to draw inferences regarding the relationships we are modeling.

References

- Allers MA, Elhorst JP (2005) Tax mimicking and yardstick competition among governments in the Netherlands. *Int Tax Publ Finance* 12(4):493–513
- Anselin L (1988) Spatial econometrics: methods and models. Kluwer, Dordrecht
- Anselin L, Le Gallo J, Jayet H (2008) Spatial panel econometrics. In: Mátyás L, Sevestre P (eds) The econometrics of panel data, fundamentals and recent developments in theory and practice, 3rd edn. Kluwer, Dordrecht, pp 627–662
- Arellano M, Bond S (1991) Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations. *Rev Econ Stud* 58(2):277–297
- Baltagi BH (2005) Econometric analysis of panel data, 3rd edn. Wiley, Chichester
- Baltagi BH, Levin D (1992) Cigarette taxation: raising revenues and reducing consumption. *Struct Chang Econ Dyn* 3(2):321–335
- Baltagi BH, Li D (2004) Prediction in the panel data model with spatial autocorrelation. In: Anselin L, Florax RJGM, Rey SJ (eds) Advances in spatial econometrics: methodology, tools, and applications. Springer, Berlin, pp 283–295
- Bhargava A, Sargan JD (1983) Estimating dynamic random effects models from panel data covering short time periods. *Econometrica* 51(6):1635–1659
- Blundell R, Bond S (1998) Initial conditions and moment restrictions in dynamic panel data models. *J Econ* 87(1):115–143
- Brueckner JK (2003) Strategic interaction among local governments: an overview of empirical studies. *Int Reg Sci Rev* 26(2):175–188
- Debarsy N, Ertur C, LeSage JP (2012) Interpreting dynamic space-time panel data models. *Stat Methodol* 9(1–2):158–171
- Elhorst JP (2003) Specification and estimation of spatial panel data models. *Int Reg Sci Rev* 26(3):244–268
- Elhorst JP (2010a) Spatial panel data models. In: Fischer MM, Getis A (eds) Handbook of applied spatial analysis. Springer, Berlin/Heidelberg/New York, pp 377–407
- Elhorst JP (2010b) Dynamic panels with endogenous interaction effects when T is small. *Reg Sci Urban Econ* 40(5):272–282
- Elhorst JP (2012) Dynamic spatial panels: models, methods and inferences. *J Geogr Syst* 14(1):5–28
- Elhorst JP, Piras G, Arbia G (2010) Growth and convergence in a multi-regional model with space-time dynamics. *Geogr Anal* 42(3):338–355
- Ertur C, Koch W (2007) Growth, technological interdependence and spatial externalities: theory and evidence. *J Appl Econ* 22(6):1033–1062
- Hsiao C, Pesaran MH, Tahmisioglu AK (2002) Maximum likelihood estimation of fixed effects dynamic panel data models covering short time periods. *J Econ* 109(1):107–150
- Kelejian HH, Prucha IR (1998) A generalized spatial two stage least squares procedure for estimating a spatial autoregressive model with autoregressive disturbances. *J Real Estate Financ Econ* 17(1):99–121
- Lee LF (2004) Asymptotic distribution of quasi-maximum likelihood estimators for spatial autoregressive models. *Econometrica* 72(6):1899–1925

- Lee LF, Yu J (2010a) Estimation of spatial autoregressive panel data models with fixed effects. *J Econ* 154(2):165–185
- Lee LF, Yu J (2010b) A spatial dynamic panel data model with both time and individual fixed effects. *Econom Theory* 26(2):564–597
- Lee LF, Yu J (2010c) Efficient GMM estimation of spatial dynamic panel data models with fixed effects. <http://www.economics.smu.edu.sg/events/Paper/LungfeiLee.pdf>
- Lee LF, Liu X, Lin X (2010) Specification and estimation of social interaction models with network structures. *Econ J* 13(2):145–176
- LeSage JP, Pace RK (2009) Introduction to spatial econometrics. CRC Press/Taylor & Francis, Boca Raton
- Parent O, LeSage JP (2010) A spatial dynamic panel model with random effects applied to commuting times. *Transp Res Part B* 44(5):633–645
- Parent O, LeSage JP (2011) A space-time filter for panel data models containing random effects. *Comput Stat Data Anal* 55(1):475–490
- Yu J, de Jong R, Lee LF (2008) Quasi-maximum likelihood estimators for spatial dynamic panel data with fixed effects when both n and T are large. *J Econ* 146(1):118–134

Christine Thomas-Agnan and James P. LeSage

Contents

83.1	Introduction to Gravity or Spatial Interaction Models	1654
83.2	Gravity or Spatial Interaction Models Based on Independence	1656
83.3	Spatial Autoregressive Interaction Models	1659
83.4	Problems That Arise in Applied Modeling of Flows	1660
83.5	Interpreting Spatial Interaction Models	1662
83.5.1	A Numerical Illustration for the Nonspatial Gravity Model	1662
83.5.2	A Numerical Illustration for the Spatial Gravity Model	1667
83.6	Conclusion	1671
	References	1672

Abstract

Spatial interaction or gravity models have been used in regional science to model flows that take many forms, for example, population migration, commodity flows, traffic flows, and knowledge flows, all of which reflect movements between origin and destination regions. This chapter focuses on spatial autoregressive extensions to the conventional least-squares gravity models that relax the assumption of independence between flows. These models, proposed by LeSage and Pace (2008, *Spatial econometric modeling of origin-destination flows*. *J Reg Sci* 48(5):941–967, 2009), define spatial dependence in this type of setting to mean that larger observed flows from an origin region A to

C. Thomas-Agnan (✉)
G.R.E.M.A.Q., Toulouse School of Economics, Toulouse, France
e-mail: christine.thomas@tse.fr.en

J.P. LeSage
Department of Finance and Economics, Texas State University – San Marcos, San Marcos,
TX, USA
e-mail: jlesage@spatial-econometrics.com

a destination region Z are accompanied by (i) larger flows from regions nearby the origin A to the destination Z , say regions B and C that are neighbors to region A , which they label origin dependence; (ii) larger flows from the origin region A to regions neighboring the destination region Z , say regions X and Y , which they label destination dependence; and (iii) larger flows from regions that are neighbors to the origin (B and C) to regions that are neighbors to the destination (X and Y), which they label origin-destination dependence. Spatial spillovers in these models can take the form of spillovers to both regions/observations neighboring the origin or destination in the dyadic relationships that characterize origin-destination flows as well as network effects that impact all other regions in the network. We set forth a simulation approach for these models that can be used to produce scalar expressions for the various types of spillover impacts that arise from changes in the explanatory variables of the model.

83.1 Introduction to Gravity or Spatial Interaction Models

Gravity models have often been used to explain origin-destination (OD) flows that arise in regional science such as trade, transportation, and migration. In the regional science literature, the gravity model has been labeled a spatial interaction model (Sen and Smith 1995) because the regional interaction is directly proportional to the product of regional size measures. In the case of interregional commodity flows, the measure of regional size is typically gross regional product or regional income. The model predicts more interaction in the form of commodity flows between regions of similar (economic) size than regions dissimilar in size. For the case of migration flows, population would be a logical measure of regional size, and in other contexts such as knowledge flows between regions, LeSage, Fischer, and Scherngell (2007) use regional knowledge stocks measured by patents to reflect size.

Theoretical motivations for spatial interaction modeling are numerous, for example, Wilson (1967) and Roy (2004) provide a macroeconomic statistical equilibrium development, Smith (1975) and Sen and Smith (1995) rely on a microeconomic choice-theoretic approach, and Fischer (2002) and Fischer and Reismann (2002) take a neural network approach that treats spatial interaction models as universal function approximations.

Historically, motivations for these models took the view that spatial interaction implies movement of entities, and that this has little to do with spatial association (Getis 1991). These models attempt to explain variation in observed flows between origin and destination regions using (i) *origin-specific* attributes that characterize the ability of the origins to generate outflows, (ii) destination-specific attributes that attract inflows, and (iii) variables reflecting the spatial separation of origin and destination regions. The traditional assumption was that including *separation variables* (such as distance, borders, language, and cultural differences between regions) should fully account for observed spatial dependence in flows. Curry (1972) was an earlier dissenter from this view, advancing a theoretical motivation for the presence of spatial dependence in flows after conditioning on conventional

variables, and Griffith and Jones (1980) reported spatial correlation in residuals of conventional spatial interaction models. The notion that use of distance functions in conventional spatial interaction models effectively captures spatial dependence in the interregional flows being analyzed was further challenged by Porojan (2001) for the case of international trade flows, and Lee and Pace (2005) for retail sales. Both studies reported residuals from conventional models that exhibited spatial dependence. Despite these findings, most applied work continued to assume independence between flow observations relying on conventional least-squares models to explain observed variation in flows. One exception is Bolduc, Laferriere, and Santarossa (1992) who explicitly model the disturbances using a spatial autoregressive process.

LeSage and Pace (2008) define spatial dependence in a spatial interaction setting to mean that larger observed flows from an origin region A to a destination region Z are accompanied by (i) larger flows from regions nearby the origin A to the destination Z , say regions B and C that are neighbors to region A , which they label *origin dependence*; (ii) larger flows from the origin region A to regions neighboring the destination region Z , say regions X and Y , which they label *destination dependence*; and (iii) larger flows from regions that are neighbors to the origin (B and C) to regions that are neighbors to the destination (X and Y), which they label *origin-destination dependence*. Using this definition of spatial dependence, modeling of spatial dependence in regional flows requires a spatial autoregressive specification.

LeSage and Pace (2008) show how to produce maximum likelihood estimates for a spatial autoregressive specification of the spatial interaction model. This model includes spatial lags of the dependent variable similar to conventional spatial autoregressive models in an effort to directly model spatial dependence in flows. Fischer and Griffith (2008) use the approach introduced by LeSage and Pace (2008) to include spatial lags for the model disturbances. LeSage and Pace (2009) show how to produce Bayesian Markov Chain Monte Carlo estimates for their spatial econometric variant of the spatial interaction model. While the motivation provided by LeSage and Pace (2008) for the spatial econometric approach to spatial interaction modeling is purely econometric, Behrens, Ertur, and Koch (2012) provide a theoretical justification for such models.

Section 83.2 introduces the conventional spatial interaction model that assumes independence between observed flows and relies on ordinary least-squares estimation methods.

In Sect. 83.3, we introduce the spatial autoregressive extension of LeSage and Pace (2008). Section 83.4 discusses a number of problems that arise in applied modeling of regional flows that can invalidate use of maximum likelihood (or Bayesian) estimation of the model from LeSage and Pace (2008). These problems provide fertile ground for future research in spatial interaction modeling.

While the focus in LeSage and Pace (2008, 2009) was on maximum likelihood and Bayesian estimation of spatial autoregressive interaction models, there is also a need to consider how estimates from these models should be properly interpreted. The subject of interpreting estimates from independent and spatial autoregressive

spatial interaction models is taken up in Sect. 83.5. We set forth a simulation approach for these models that can be used to produce scalar expressions for the various types of spillover impacts that arise from changes in the explanatory variables of the model. Spatial spillovers in these models can take the form of spillovers to both regions/observations neighboring the origin or destination in the dyadic flow relationships that characterize origin-destination flows as well as network effects that impact all other regions in the network. We also make the point that interpretation of estimates from conventional independent spatial interaction models may be improved using this approach.

83.2 Gravity or Spatial Interaction Models Based on Independence

Regression models attempt to explain variation in the n^2 flows between the n regions in a closed network of regional flows. The $n \times n$ flow matrix Y is converted to an $n^2 \times 1$ vector by stacking columns. The flow matrix might be arranged so the i, j th element reflects a flow from region j to region i , which has been labeled an origin-centric flow arrangement by LeSage and Pace (2008). Many trade models rely on the convention that the i, j th element of the flow matrix represents a flow from region i to j , which would be a destination-centric arrangement of the flows. If we let y^o denote the origin-centric vector of flows and y^d a vector created by stacking columns from a destination-centric arrangement, there is a vec-permutation matrix P that can be used to relate these two different orderings. Specifically, $Py^o = y^d$, and using properties of permutation matrices, $y^o = P^{-1}y^d = P'y^d$.

A regression model that has been labeled a gravity model captures the notion that the size of the two regions and the distance between them are important factors that determine the magnitude of flows between regions. For example, if one starts with the standard gravity model (c.f., Eq. (6.4) in Sen and Smith 1995) shown in Eq. (83.1) and applies a log transformation, the regression in Eq. (83.2) arises.

$$\mu(i,j) = CX_o(i)X_d(j)H(i,j) \quad (83.1)$$

In Eq. (83.1), $\mu(i,j)$ represents the expected flows from region i to region j (assuming a destination-centric flow matrix), while $X_d(i), X_o(j)$ denote sizes of the destination and origin and $G(i,j)$ represents resistance or deterrence to flows between the origin and destination, typically modeled using some function of distance between regions i and j . To facilitate the log transformation, $X_o(i)$ can be specified using $X_o(i)^{\beta_o}$ and similarly, $X_d(j) = X_d(j)^{\beta_d}$, while $H(i,j)$ is some function of distance between regions i and j , for which we might use a power function, $D(i,j)^\gamma$, where $D(i,j)$ is the distance between regions i and j .

A point made by LeSage and Pace (2009) is that conventional work with these models has relied on mathematics emphasizing dyads i, j which has severe limitations for thinking about flows in the context of a network. Spatial dependence

reflects relationships between observations, and is typically modeled using vectors and spatial weight matrices to express relations between observations. LeSage and Pace (2008) use the matrix/vector representation of the log-transformed dyad expression in Eq. (83.1) shown in Eq. (83.2), which more closely mirrors notation from conventional regression modeling. It should also be noted that another population formulation directly models flows as: $F(i,j) = \exp\left(\sum_{k=1}^R \beta_k \log X_{kij}\right) + \varepsilon_{ij}$, where the disturbance term is additive. This produces a Poisson model suitable for flows taking the form of counts and requires maximum likelihood estimation (Gourieroux et al. 1984).

$$y = \alpha \iota_{n^2} + X_o \beta_o + X_d \beta_d + \gamma g + \varepsilon \quad (83.2)$$

In Eq. (83.2), y is an $n^2 \times 1$ vector of (logged) flows constructed by stacking the columns of the $n \times n$ flow matrix Y , where we will assume a destination-centric organization throughout this chapter. Similarly, applying the log transformation to the $n \times n$ matrix of distances $D(i,j)$ between the n destination and origin regions and stacking the columns results in a vector of logged distances g , with associated coefficient γ . LeSage and Pace (2008) show that $X_o = \iota_n \otimes X$, where X is an $n \times R$ matrix of characteristics for the n regions, \otimes represents a Kronecker product, and ι_n is an $n \times 1$ vector of ones. In the simplest case, X might represent a vector with the appropriate size measure for each region, but without loss of generality this may be a matrix containing R characteristics of the regions that are thought to explain variation in flows. We note that this represents a general case where the same set of explanatory variables is used for both origins and destinations. A special case might involve selection of a subset of variables in the matrix X for use as origin characteristics, and another subset of variables for the destination characteristics. However, the general case maybe the preferred approach to specification, since inclusion of additional unimportant explanatory variables does not bias least-squares estimates, whereas exclusion of important explanatory variables can result in omitted variables bias.

The Kronecker product repeats the same values of the n regions in a strategic fashion to create a vector (or matrix) of sizes associated with each origin region, hence use of the notation X_o to represent these explanatory variables reflecting origin characteristics. Ultimately, use of Kronecker products in conjunction with matrix algebra allowed LeSage and Pace (2008) to express simple estimators that avoid storing multiple copies of the same numerical values, which is computationally inefficient. The matrix/vector $X_d = X \otimes \iota_n$ arranges the n regional characteristics to match the vector y , producing explanatory variables associated with each destination region. The vectors β_o and β_d are $R \times 1$ parameter vectors associated with the origin and destination region characteristics, respectively. The scalar parameter γ reflects the effect of the vector of logged distances g on flows, which is traditionally thought to be negative. The parameter α denotes the constant term parameter, and the $n^2 \times 1$ vector ε represents zero mean, constant variance, zero covariance disturbances, consistent with the Gauss-Markov least-squares assumptions. We note that the assumption of normally distributed disturbances consistent with least-squares implies that the dependent variable flows are also normally

distributed. This is not consistent with some flows which represent count data, for example, counts of persons migrating or commuters traveling from one region to another. However, the log transformation may help to produce more normally distributed flows. We will have more to say about this issue in Sect. 83.4, where problems that affect maximum likelihood estimation of spatial interaction models are discussed.

LeSage and Pace (2008) note that the algebra of Kronecker products can be used to avoid the need to form $n^2 \times R$ matrices X_o, X_d which require a great deal of computer storage involving repeated numerical values. This can be seen by examining the $(2R + 2) \times (2R + 2)$ moment matrix formed using: $Z = (I_{n^2} \ X_o \ X_d \ g)$ shown in Eq. (83.3), where we use G to represent the $n \times n$ matrix of logged distances.

$$Z'Z = \begin{pmatrix} n^2 & 0_k & 0_k & I_n'G I_n \\ 0_k' & nX'X & 0_k'0_k & X'G I_n \\ 0_k & 0_k'0_k & nX'X & X'G I_n \\ I_n'G I_n & I_n'G'X & I_n'G'X & \text{tr}(G^2) \end{pmatrix} \quad (83.3)$$

Similarly, the matrix product $Z'y$ involving the matrix Z' of dimension $(2R + 2) \times n^2$ and the $n^2 \times 1$ vector of flows can be formed as shown in Eq. (83.4), where tr denotes the trace operator.

$$Z'\text{vec}(Y) = \begin{pmatrix} I_n^2 \\ X_o' \\ X_d' \\ g' \end{pmatrix} \quad y = \begin{pmatrix} I_n'Y I_n \\ X'Y I_n \\ X'Y' I_n \\ \text{tr}(GY) \end{pmatrix} \quad (83.4)$$

This allows calculation of the parameter estimates $\delta = (\alpha \ \beta_o \ \beta_d \ \gamma)'$ using only the $n \times R$ matrix X , the $n \times n$ flow matrix Y , and the $n \times n$ matrix of logged distances G that appear in $Z'Z$ and $Z'y$, as shown in Eq. (83.5).

$$\hat{\delta} = [(1/n^2)Z'Z]^{-1}(1/n^2)Z'y \quad (83.5)$$

Interpretation of the estimates β_o, β_d has followed that used in typical regression, where these parameters reflect the influence (positive or negative) of changes in origin and destination characteristics on the magnitude of flows. Since the model has been log-transformed, these estimates can be interpreted as elasticities. A negative estimate for the r th destination characteristic indicates that this reduces flows to the destination, whereas a positive coefficient points to a factor that increases flows to the destination. A similar interpretation applies to the coefficients β_o , which measure the positive or negative influence of origin characteristics on flows. We will have more to say about this approach to interpreting the coefficients β_o, β_d later. The coefficient γ should be negative, consistent with the notion that (logged) distance acts as a friction to reduce flows.

83.3 Spatial Autoregressive Interaction Models

Intuitively, changes to the characteristics of a single region i will impact both inflows and outflows to all other regions engaged or connected with region i as either an origin or destination. For example, a (*ceteris paribus*) decrease in taxes in region i would lead to inflows of population to this region from (potentially) all other regions and a decrease in outflows of population to (potentially) all other regions.

LeSage and Pace (2008) suggest that flows across networks involving origins and destinations are likely to exhibit spatial dependence. They define spatial dependence in this type of setting to mean that larger observed flows from an origin region A to a destination region Z are accompanied by (i) larger flows from regions nearby the origin A to the destination Z , say regions B and C that are neighbors to region A , which they label origin dependence; (ii) larger flows from the origin region A to regions neighboring the destination region Z , say regions X and Y , which they label destination dependence; and (iii) larger flows from regions that are neighbors to the origin (B and C) to regions that are neighbors to the destination (X and Y), which they label origin-destination dependence.

Casual observation of migration flows in a network of counties is consistent with this type of observation. If there are a large number of migrants moving away from a county A (say a county near the Detroit metropolitan area), we would expect to see migrants also moving away from other counties B and C near Detroit (presumably due to unfavorable labor market conditions). Similarly, if a large number of migrants are moving into a county Z (say a county in the Austin metropolitan area), we would expect to see migrants also moving into other counties X and Y in the Austin metropolitan area (presumably because of favorable labor market conditions).

LeSage and Pace (2008, 2009) propose a spatial regression extension of the independent empirical gravity model from Eq. (83.2) shown in Eq. (83.6).

$$Ay = \alpha I_{n^2} + X_o \beta_o + X_d \beta_d + g\gamma + \varepsilon \quad (83.6)$$

$$\begin{aligned} A &= (I_{n^2} - \rho_o W_o)(I_{n^2} - \rho_d W_d) \\ &= (I_{n^2} - \rho_o W_o - \rho_d W_d + \rho_w W_w) \end{aligned}$$

$$W_o = I_n \otimes W$$

$$W_d = W \otimes I_n$$

$$W_w = W_o \otimes W_d = W_d \otimes W_o = W \otimes W$$

The term A can be viewed as a spatial filter that captures origin-based dependence, destination-based dependence, and origin-destination-based dependence. (The filter implies a restriction that $\rho_w = -\rho_o \rho_d$. This restriction need not be imposed during estimation, so we address the more general case here and allow for an unrestricted parameter ρ_w .) The model and associated data generating

process (DGP) for the spatial autoregressive interaction model take the forms shown in Eqs. (83.7) and (83.8), respectively, where we rely on the earlier definitions of Z and δ .

$$y = \rho_o W_o y + \rho_d W_d y + \rho_w W_w y + Z\delta + \varepsilon \quad (83.7)$$

$$y = (I_{n^2} - \rho_o W_o - \rho_d W_d + \rho_w W_w)^{-1} (Z\delta + \varepsilon) \quad (83.8)$$

The spatial lag formed by the matrix product $W_d y$ extracts flows from neighbors to each destination region in the vector of origin-destination flow dyads to form a linear combination of flows from neighboring destinations. In the case where the $n \times n$ spatial weight matrix W represents a fixed number, say m , of equally weighted nearest neighbors, the spatial lag vector would contain an average of flows from the m neighboring destinations. The matrix W is a conventional (row-normalized) spatial weight matrix of the type used in cross-sectional regressions involving n regions. This spatial lag captures destination-based dependence, with the parameter ρ_d measuring the strength of destination-based dependence.

A similar interpretation applies to the spatial lag formed by the product $W_o y$, which reflects a linear combination of flows from regions neighboring the origin, again for each origin-destination dyad in the flow vector. The scalar parameter ρ_o reflects the strength of origin-based dependence. The spatial lag $W_w y$ forms a linear combination of flows from neighbors to the origin and flows from neighbors to the destination, and the parameter ρ_w represents the magnitude of this type of dependence.

The stability restrictions for the spatial dependence parameters require that $1/\lambda_{\min} < \rho_o + \rho_d + \rho_w < 1$, where λ_{\min} is the minimum eigenvalue of the matrix W . In practice, values of -1 are often used to replace $1/\lambda_{\min}$, since this avoids the need to calculate the minimum eigenvalue of the matrix W .

LeSage and Pace (2008) provide details concerning maximum likelihood estimation for the spatial autoregressive interaction model, and LeSage and Pace (2009) set forth a Bayesian MCMC estimation scheme. Both of these exploit the computationally efficient moment matrices involving the sample data expressed using the smaller dimension matrices.

83.4 Problems That Arise in Applied Modeling of Flows

Maximum likelihood estimation methods require that the disturbances in the model follow a normal distribution, which implies that the dependent variable flows are also normally distributed. As already noted, many flows are more properly viewed as count data magnitudes, for example, flows of population or commuters migrating or traveling between regions. There are limitations to the ability of the log transformation to convert count data to a form consistent with a normal

distribution, especially when a large number of flows between regions take on zero values. For a flow matrix involving small regions, there are likely to be a large number of zero flow magnitudes. For example, migration flows between US counties in the contiguous states over a 5-year period exhibit zeros for over 90 % of the county-to-county flows. A solution for problems involving large numbers of zero flows as well as small flows would be development of Poisson variants of the spatial autoregressive spatial interaction model. Some work has been done in this area. Lambert, Brown, and Florax (2010) set forth a two-stage estimation procedure for a spatial autoregressive Poisson model, that is one not representing a spatial interaction model. LeSage, Fischer, and Scherngell (2007) introduce spatially structured origin and destination effects in a Poisson model involving counts of interregional patent citations, but their model does not involve spatial lags of the dependent variable. LeSage and Llano (2006) introduce a Tobit variant of a model that contains spatially structured origin and destination effects parameters, which can address cases involving smaller numbers of zero flows. Ranjan and Tobias (2007) also use a Tobit approach but rely on semi-parametric origin and destination effects parameters. The use of Tobit models is an attempt to address a common practice where practitioners modify the dependent variable vector using: $\ln(1 + y)$ to accommodate the log transformation. Since this transformation ignores the mixed discrete/continuous nature of the flow distribution, it should lead to downward bias in the coefficient estimates for the model. An appropriate approach to addressing the problem of a large number of zero flows as well as small flows and the count nature of many flows remains an area for future research.

Another factor contributing to non-normality in flow magnitudes is the presence of large flows within regions, those located on the main diagonal of the flow matrix, relative to smaller flows between regions, those on the off-diagonal elements. Since the objective of spatial interaction modeling is typically a model that explains variation in interregional flows, practitioners often view intraregional flows as a nuisance. Some common practices are (i) to simply set observed intraregional flows to zero values (Tiefelsdorf 2003; Fischer et al. 2006) and (ii) introduce dummy variables for these observations (Behrens et al. 2012). For the case of the independence model, these approaches are fine, but they can have adverse impacts on spatial autoregressive interaction models. Inclusion of zero magnitudes for intraregional flows in a model that includes spatial lags such as W_{oy} , W_{dy} will produce aberrant observations when these flows become part of the linear combination of neighboring values to the origin or destination.

LeSage and Pace (2008) propose using a separate set of explanatory variables in the spatial autoregressive interaction specification to deal with large flow magnitudes on the main diagonal of the flow matrix. This separate model is embedded into the specification by adjusting the explanatory variables matrices X_o , X_d and the intercept vector ι_n to have zero values for the n observations associated with the main diagonal elements (intraregional flows) of the flow matrix. They then introduce an additional explanatory variables matrix containing only n nonzero observations, those associated with intraregional flows that were set to zero in the

matrices X_o, X_d . In addition, new intercept vectors are introduced: one that contains zeros for observations associated with intraregional flows and ones for all others, and a second that contains ones for only the intraregional flow observations.

This approach allows nonzero intraregional flows to be included in the dependent variable vector y which is used to form the spatial lags W_{oy}, W_{dy}, W_{wy} . The part of flow variation associated with the large diagonal elements is explained by the embedded model variables allowing the coefficient estimates associated with the adjusted explanatory variables to more accurately characterize variation in interregional flows. LeSage and Fischer (2010) provide an example of the improvement that arises from this approach.

Assuming the problems of zero flows and the count nature of some flow magnitudes can be solved for the case of the spatial autoregressive spatial interaction model, there is still the issue of how to properly interpret estimates from this model.

83.5 Interpreting Spatial Interaction Models

A first point to note is that we should not interpret the coefficient estimates β_d, β_o and γ as if they were least-squares estimates that reflect partial derivative changes in the dependent variable associated with changes in the explanatory variables. LeSage and Pace (2009) point out that this mistaken approach to spatial autoregressive (SAR) models has been used in much of the past spatial econometrics literature.

We present a method that can be used to relate changes in characteristics of a single region i to flows across the $n \times n$ network of flows between the n regions for the case of the spatial autoregressive interaction model. This issue has not been tackled in the literature, yet it is essential for interpreting the coefficient estimates β_o, β_d in the spatial autoregressive interaction model.

83.5.1 A Numerical Illustration for the Nonspatial Gravity Model

Prior to setting forth our method for quantifying how changes in the r th characteristic of region i impact flows, we provide a simple numerical illustration to fix ideas. Using the DGP in Eq. (83.8), we generated a set of flows using $n = 8$ regions with $\beta_d = 1, \beta_o = 0.5, \delta = -0.5, \rho_d = 0.4, \rho_o = 0.4$, and $\rho_w = -\rho_o \times \rho_d = -0.16$. No disturbance term was used, and the single vector $x' = (40\ 30\ 20\ 10\ 7\ 10\ 15\ 25)$ was used, so we have the case where $R = 1$. A set of n latitude and longitude coordinates (both equal to $1, 2, \dots, 8$) were used to produce an n^2 vector of (logged) distances g and the associated spatial weight matrix W based on two nearest (distanced) neighbors. The systematic order of the latitude-longitude coordinates produces regions configured to lie on a line, with a simplified spatial weight matrix configuration. For example, region 3 has regions 2 and 4 as the two nearest neighbors, region 4 has

regions 3 and 5 as the two neighbors, and so on. This greatly simplifies things relative to real-world data. The weight matrix for our example is shown in Eq. (83.9).

$$W = \begin{pmatrix} 0.0 & 0.5 & 0.5 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 \\ 0.5 & 0.0 & 0.5 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.5 & 0.0 & 0.5 & 0.0 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.5 & 0.0 & 0.5 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.5 & 0.0 & 0.5 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.0 & 0.5 & 0.0 & 0.5 & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.5 & 0.0 & 0.5 \\ 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.5 & 0.5 & 0.0 \end{pmatrix} \quad (83.9)$$

A discrete change in each element/region $i = 1, \dots, 8$ of the vector x of one unit was made and the discrete changes arising in the $n \times n$ flow matrix as a result of these perturbations were recorded. For each change in the value x_i for a single region i , a new flow matrix was generated and subtracted from the original flow matrix to illustrate how changes in the characteristics of a single region impact the matrix of flows.

An important point to note here is that unlike the conventional spatial autoregressive model where the matrix X contains characteristics for each of the n regions, the matrices X_o , X_d in the spatial autoregressive interaction model strategically repeats values of the $n \times R$ matrix X to form $n^2 \times R$ matrices $X_o = I_n \otimes X$, $X_d = X \otimes I_n$. An implication is that when we change the characteristic/element of a single region i (which we denote using x_i), it produces a set of changes in n elements of the matrix X_o and changes in n elements of the matrix X_d . Together, this set of $2n$ altered values in the matrices X_o , X_d produce the change in flows that results from changing characteristics of the i th region, that is $x_i + 1$. This has computational implications for how we calculate the effects arising from changes in the explanatory variables of this model. Unlike the conventional SAR model, we do not need to calculate changes in each of the n^2 elements of the vectors X_o and X_d to produce scalar summary measures of the impact of these changes on the flows. Although this approach is valid, it requires more computational effort. Instead we can consider only n changes in each observation i of the matrix/vector X as producing a total derivative response. There will be a vector of $n^2 \times 1$ responses in the flows (which can be viewed as a change in the $n \times n$ flows matrix Y) arising from a change in a single characteristic of the i th region, x_i . This single element total derivative change works through a series of $2n$ associated changes that arise in the $n^2 \times R$ model explanatory variables X_o , X_d .

Intuitively, increasing a single region i 's characteristic (say the size of region x_i) means this region will (i) attract more inflows as a destination from all n regions (including itself which takes the form of more intraregional flows within region i) and (ii) produce increased outflows to all n regions (including itself). This facet of changes in the characteristic of a single region is what accounts for the model repeating the same altered value of x_i (the new size for region i) n times in the vector/matrix X_o , and n times in X_d . Given this, it is computationally inefficient to consider conventional partial derivatives that would independently change each of

the n^2 elements in X_o or X_d and examine their impact on the flow matrix. Changes to individual elements of X_o and X_d need not be considered given the structure of the model (and associated data generating process).

There may be applied modeling situations where different explanatory variables are used to model the origin and destination characteristics of the regions that are thought to be important for explaining variation in flows. In these situations, the argument used above regarding changes to a single explanatory variable x_i for each region will not be valid. The more computationally inefficient approach of using conventional partial derivatives that would independently change each of the n^2 elements in X_o and X_d would need to be used in order to examine the impact of these changes on the flow matrix. We discuss this type of situation when providing a numerical illustration.

Results showing the changes in the $n \times n$ flow matrix associated with a change in the third region's characteristic, x_3 , by one unit for the case of the independent (nonspatial) gravity model in Eq. (83.2) are shown in Eq. (83.10). These were produced by setting $\rho_o = \rho_d = \rho_w = 0$ in the spatial gravity model from Eq. (83.8), which results in the independent gravity model from Eq. (83.2).

$$\Delta Y / \Delta x_3 = \begin{pmatrix} 0.00 & 0.00 & 1.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 1.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.50 & 0.50 & 1.50 & 0.50 & 0.50 & 0.50 & 0.50 & 0.50 \\ 0.00 & 0.00 & 1.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 1.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 1.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 1.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 1.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \end{pmatrix} \quad (83.10)$$

The role of the independence assumption is clear in Eq. (83.10), where we see from row 3 that the change of outflows from region 3 to all other regions equals 0.5, which is the value of the coefficient β_o in our example. Similarly, column 3 exhibits identical changes in inflows to region 3, taking the value 1 of the coefficient β_d in our example. The diagonal (3,3) element reflects a response equal to $\beta_o + \beta_d$, the sum of the changes in flows into and out of region 3, reflecting the change in intraregional flows arising from the change in x_3 . We have only $2n$ nonzero changes in flows by virtue of the independence assumption. All changes involving flows into and out of regions other than region 3 are zero.

Our method for producing scalar summary measures of the impacts arising from changes in characteristics of the regions involves averaging over the cumulative flow impacts associated with changes in all regions, $i = 1, \dots, n$, analogous to the approach taken by LeSage and Pace (2009) for the conventional SAR model. Doing this produces $\sum_{i=1}^n (\Delta Y / \Delta x_i)$, a cumulative total effects (TE) matrix shown in Eq. (83.11), which is the sum of $n = 8$ different changed flow matrices of the type shown in Eq. (83.10) for the case where $i = 3$. This matrix (TE) can be decomposed into flow matrices reflecting origin effects (OE), destination effects (DE), network effects (NE), and intraregional effects (IE) arising from changing a single characteristic in all regions by one unit.

$$TE = \begin{pmatrix} 1.5 & 1.5 & 1.5 & 1.5 & 1.5 & 1.5 & 1.5 & 1.5 \\ 1.5 & 1.5 & 1.5 & 1.5 & 1.5 & 1.5 & 1.5 & 1.5 \\ 1.5 & 1.5 & 1.5 & 1.5 & 1.5 & 1.5 & 1.5 & 1.5 \\ 1.5 & 1.5 & 1.5 & 1.5 & 1.5 & 1.5 & 1.5 & 1.5 \\ 1.5 & 1.5 & 1.5 & 1.5 & 1.5 & 1.5 & 1.5 & 1.5 \\ 1.5 & 1.5 & 1.5 & 1.5 & 1.5 & 1.5 & 1.5 & 1.5 \\ 1.5 & 1.5 & 1.5 & 1.5 & 1.5 & 1.5 & 1.5 & 1.5 \\ 1.5 & 1.5 & 1.5 & 1.5 & 1.5 & 1.5 & 1.5 & 1.5 \end{pmatrix} \quad (83.11)$$

The matrix of cumulative intraregional effects can be constructed using the main diagonal elements of the TE matrix, $IE(i, i) = \sum_{i=1}^n (\Delta Y_{(i,i)} / \Delta x_i)$. The matrix of (cumulative) intraregional effects is shown in Eq. (83.12), where we see that these are identical and equal to the value of the coefficients $\beta_o + \beta_d$ from our example. These are located on the main diagonal which reflects changes in intraregional flows.

$$IE = \begin{pmatrix} 1.5 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 \\ 0.0 & 1.5 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 1.5 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.0 & 1.5 & 0.0 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.0 & 1.5 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 1.5 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 1.5 & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 1.5 \end{pmatrix} \quad (83.12)$$

The matrix of (cumulative) origin effects can be constructed using the i th row of the flow changes matrix excluding the intraregional effect from the diagonal element. Specifically, $OE(i, .) = \sum_{i=1}^n (\Delta Y_{(i,.)} / \Delta x_i) - IE(i, i)$, where we use $OE(i, .)$ and $\Delta Y_{(i,.)}$ to denote the i th row of the OE matrix and flow changes matrix ΔY . The result is shown in Eq. (83.13), where we see that these are identical and equal to the value of the coefficient β_o from our example. The main diagonal is zero since this reflects changes in intraregional flows which we have excluded from our definition of origin effects.

$$OE = \begin{pmatrix} 0.0 & 0.5 & 0.5 & 0.5 & 0.5 & 0.5 & 0.5 & 0.5 \\ 0.5 & 0.0 & 0.5 & 0.5 & 0.5 & 0.5 & 0.5 & 0.5 \\ 0.5 & 0.5 & 0.0 & 0.5 & 0.5 & 0.5 & 0.5 & 0.5 \\ 0.5 & 0.5 & 0.5 & 0.0 & 0.5 & 0.5 & 0.5 & 0.5 \\ 0.5 & 0.5 & 0.5 & 0.5 & 0.0 & 0.5 & 0.5 & 0.5 \\ 0.5 & 0.5 & 0.5 & 0.5 & 0.5 & 0.0 & 0.5 & 0.5 \\ 0.5 & 0.5 & 0.5 & 0.5 & 0.5 & 0.5 & 0.0 & 0.5 \\ 0.5 & 0.5 & 0.5 & 0.5 & 0.5 & 0.5 & 0.5 & 0.0 \end{pmatrix} \quad (83.13)$$

The matrix DE of (cumulative) destination effects is based on using the i th column of the flow changes matrix, excluding the intraregional effect from the diagonal element. Of course, the OE and DE definitions would reverse if we were

relying on an origin-centric flow matrix instead of the destination-centric one. Specifically, $DE(., i) = \sum_{i=1}^n (\Delta Y_{(.,i)} / \Delta x_i) - IE(i, i)$, where we use $DE(., i)$ and $\Delta Y_{(.,i)}$ to denote the i th column of the DE matrix and flow changes matrix ΔY . The result is shown in Eq. (83.14), where we see that these are identical and equal to the value of the coefficient β_d from our example. Again, the main diagonal is zero since this reflects changes in intraregional flows which we have excluded from our definition of destination effects.

$$DE = \begin{pmatrix} 0.0 & 1.0 & 1.0 & 1.0 & 1.0 & 1.0 & 1.0 & 1.0 \\ 1.0 & 0.0 & 1.0 & 1.0 & 1.0 & 1.0 & 1.0 & 1.0 \\ 1.0 & 1.0 & 0.0 & 1.0 & 1.0 & 1.0 & 1.0 & 1.0 \\ 1.0 & 1.0 & 1.0 & 0.0 & 1.0 & 1.0 & 1.0 & 1.0 \\ 1.0 & 1.0 & 1.0 & 1.0 & 0.0 & 1.0 & 1.0 & 1.0 \\ 1.0 & 1.0 & 1.0 & 1.0 & 1.0 & 0.0 & 1.0 & 1.0 \\ 1.0 & 1.0 & 1.0 & 1.0 & 1.0 & 1.0 & 0.0 & 1.0 \\ 1.0 & 1.0 & 1.0 & 1.0 & 1.0 & 1.0 & 1.0 & 0.0 \end{pmatrix} \quad (83.14)$$

The matrix of cumulative network effects represents all flow changes that spill-over on regions other than the origin and destination region whose characteristics were changed. This can be constructed by subtraction: $TE - IE - OE - DE = NE$. For the nonspatial gravity model, the cumulative network effects matrix NE contains all zeros, since this model does not allow for spillovers to regions not involving origin and destination regions by virtue of the independence assumption. Scalar summary measures of the total effects as well as the decomposition into origin, destination, intraregional, and network effects can be constructed using averages of the (matrices) of cumulated changes in flows. This is accomplished by averaging over row-sums and then column-sums, which follows the approach taken by LeSage and Pace (2009) for the SAR model. This produces the results shown in the first column of Table 83.1. (One can also average over column-sums and then row-sums to produce identical results as noted by LeSage and Pace (2009).)

Applying our decomposition with this computationally inefficient approach would lead to scalar summary measures for the impact of changing all elements in the vector X_o presented in the second column of Table 83.1. Similarly, our decomposition with this approach would lead to scalar summary measures for the impact of (independently) changing each element in the vector X_d shown in the third column of Table 83.1. The sum of these two sets of scalar summary effects estimates constructed using independent changes in all elements of X_o and X_d shown in the fourth column of Table 83.1 equals the result shown in the first column. We will see that this is also the case for the spatial autoregressive variant of the gravity model.

An important point to note is that this approach differs from the conventional interpretation of nonspatial gravity models where the coefficient β_o is interpreted as a partial derivative reflecting the impact of changes in origin characteristics and β_d that is associated with changing destination characteristics. Although the conventional approach that used the coefficient sum $\beta_o + \beta_d$ as a measure of the total effect on flows arising from changes in origin and destination characteristics would

Table 83.1 Scalar summary measures of effects for the nonspatial model from a change in the (single) characteristic x averaged over all regions $i = 1, \dots, 8$

	Δx_i	$\Delta X_{o,i}$	$\Delta X_{d,i}$	$\Delta X_{o,i} + \Delta X_{d,i}$
Origin effects	0.4375	0.4375	0.0000	0.4375
Destination effects	0.8750	0.0000	0.8750	0.8750
Intraregional effects	0.1875	0.0625	0.1250	0.1875
Network effects	0.0000	0.0000	0.0000	0.0000
Total effects	1.5000	0.5000	1.0000	1.5000

produce a correct inference, the appropriate decomposition into origin, destination, and intraregional effects has been missing from this literature.

Another point is that one can use changes in each element of the $n^2 \times 1$ vectors X_o and X_d to arrive at the same scalar summary measures as shown in Table 83.1. However, this would require that we sequence through changes in n^2 individual elements of X_o and also n^2 elements of X_d , recording the change in the $n \times n$ matrix of flows that arise from this sequence of $2n^2$ changes, which is computationally much more difficult. We would also need to aggregate the changes in flows arising from changes in both X_o and X_d to produce final results. To avoid this, we can exploit the special structure of the $n^2 \times R$ matrices X_o, X_d as they relate to the underlying $n \times R$ matrix X .

As noted above, there could be applied modeling situations where practitioners choose to include a specific characteristic only in the X_o or X_d vector, but not in both. As an example, consider a model for commuting-to-work flows. The number of residents might be used as a size measure for origin regions whereas the number of business establishments might be used as a size measure for the destination regions. In this case, it might be more appropriate for interpretative purposes to report separately scalar effects summaries arising from the calculations involving changing all elements in the vector X_o and X_d . We will have more to say about this in the next section.

83.5.2 A Numerical Illustration for the Spatial Gravity Model

Using the same numerical values set forth in the previous section, but setting $\rho_o = \rho_d = 0.4$ and $\rho_w = -\rho_o\rho_d = -0.16$, we carried out the same experiment where each value of $x_i, i = 1, \dots, 8$ was changed by one unit. The resulting changes in the flow matrix were recorded, with the total flow effects arising from the change in x_3 shown in Eq. (83.15).

$$\Delta Y / \Delta x_3 = \begin{pmatrix} 0.688 & 0.688 & 2.064 & 0.612 & 0.309 & 0.246 & 0.233 & 0.233 \\ 0.688 & 0.688 & 2.064 & 0.612 & 0.309 & 0.246 & 0.233 & 0.233 \\ 1.376 & 1.376 & 2.752 & 1.300 & 0.997 & 0.934 & 0.921 & 0.921 \\ 0.650 & 0.650 & 2.026 & 0.574 & 0.271 & 0.208 & 0.195 & 0.195 \\ 0.498 & 0.498 & 1.875 & 0.423 & 0.119 & 0.056 & 0.044 & 0.044 \\ 0.467 & 0.467 & 1.843 & 0.391 & 0.088 & 0.025 & 0.012 & 0.012 \\ 0.460 & 0.460 & 1.837 & 0.385 & 0.082 & 0.018 & 0.006 & 0.006 \\ 0.460 & 0.460 & 1.837 & 0.385 & 0.082 & 0.018 & 0.006 & 0.006 \end{pmatrix} \quad (83.15)$$

One difference between this spatial model result and the nonspatial model is the presence of network effects, shown by the nonzero elements in rows and columns other than 3. This means that a change in say the attractiveness of region 3 impacts flows throughout the network. Of course, the largest impacts reside in the 3rd row and column, since the change in attractiveness of region 3 has the largest impact on flows into and out of region 3 from all other regions. The magnitude of impact declines as we move further from the (3,3) element in the up/down or left/right direction in column and row 3. This arises due to decay with higher-order neighbors typical of spatial autoregressive processes. We see a similar pattern for elements not in the third row and column, where the change in flows decline in magnitude for elements further away from the (3,3) element. This reflects a decline in the magnitude of network spillovers with an increase in the number of paths through which the flows must pass.

It is also important to note that the interpretation of partial derivatives in cross-sectional spatial models such as this is that these reflect a long-run, steady-state equilibrium. The estimated changes in flows would be those that arise in response to the increased attractiveness of region 3 as we move to a new steady-state equilibrium. For example, we would conclude that changes in the attractiveness of region 3 would produce these changes in flows throughout the network, reflecting the level of flows we would expect to see in a new steady-state equilibrium.

Applying our approach for calculating scalar summary measures of the impacts arising from changes in characteristics of the regions described in the previous section we arrive at $TE = \sum_{i=1}^n (\Delta Y / \Delta x_i)$, shown in Eq. (83.16). The most obvious facet of the cumulative TE matrix is that the effects are much larger than in the case of the nonspatial gravity model. An examination of the components' (IE, OE, DE, NE) decomposition shows the source of these differences in effects on flows arising from changes in regional characteristics. A similarity with the nonspatial model TE matrix is that total effects are identical for all observations/regions, which is always the case for SAR models. This is because the spatial weight matrix W has row-sums of unity (see Elhorst 2010).

$$TE = \begin{pmatrix} 4.166 & 4.166 & 4.166 & 4.166 & 4.166 & 4.166 & 4.166 & 4.166 \\ 4.166 & 4.166 & 4.166 & 4.166 & 4.166 & 4.166 & 4.166 & 4.166 \\ 4.166 & 4.166 & 4.166 & 4.166 & 4.166 & 4.166 & 4.166 & 4.166 \\ 4.166 & 4.166 & 4.166 & 4.166 & 4.166 & 4.166 & 4.166 & 4.166 \\ 4.166 & 4.166 & 4.166 & 4.166 & 4.166 & 4.166 & 4.166 & 4.166 \\ 4.166 & 4.166 & 4.166 & 4.166 & 4.166 & 4.166 & 4.166 & 4.166 \\ 4.166 & 4.166 & 4.166 & 4.166 & 4.166 & 4.166 & 4.166 & 4.166 \\ 4.166 & 4.166 & 4.166 & 4.166 & 4.166 & 4.166 & 4.166 & 4.166 \end{pmatrix} \quad (83.16)$$

The cumulative intraregional effects matrix is shown in Eq. (83.17), where we see that the values are not equal to $\beta_o + \beta_d$ as in the nonspatial model. They are also not equal to the diagonal elements from the cumulative TE matrix. This is because there are feedback loops that arise in spatial models, where impacts on neighbors work their way back to the own region. To see this, consider that spatial

autoregressive models rely on a data generating process: $y = (I_n - \rho W)^{-1}(X\beta + \varepsilon)$, where the matrix inverse can be expressed as an infinite series: $I_n + \rho W + \rho^2 W^2 + \rho^3 W^3 + \dots$. The matrix W has zeros on the main diagonal, but the matrices W^2, W^3, \dots do not. This is because by virtue of the definition of a second-order neighbor reflected by the matrix W^2 , region i is a second-order neighbor to itself, a neighbor to a neighboring region. The feedback effects on intraregional flows account for some of the difference between the value of 4.166 for the main diagonal of the TE matrix in the spatial model and the nonspatial model, where we found a value of 1.5.

$$IE = \begin{pmatrix} 2.632 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 \\ 0.000 & 2.747 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 \\ 0.000 & 0.000 & 2.752 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 \\ 0.000 & 0.000 & 0.000 & 2.728 & 0.000 & 0.000 & 0.000 & 0.000 \\ 0.000 & 0.000 & 0.000 & 0.000 & 2.728 & 0.000 & 0.000 & 0.000 \\ 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 2.752 & 0.000 & 0.000 \\ 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 2.747 & 0.000 \\ 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 2.632 \end{pmatrix} \quad (83.17)$$

The nonzero network effects (NE) account for the remaining differences, as can be seen from the diagonal of the matrix for these cumulative effects, shown in Eq. (83.18). (The values from the diagonal of NE and IE do not exactly equal those of TE because digits were truncated when forming the matrices for presentation.) Nonzero network effects also feedback onto intraregional flows, and these account for a large part of the difference between the spatial and nonspatial model effects estimates. An important point to keep in mind is that variation in all of the effects estimates over the regions might be greater than in our simple example, where the spatial configuration of the regions represents one of the simplest. The magnitudes of network effects will depend on the spatial configuration of the regions involved, with regions that have more links to other regions experiencing larger network effects relative to regions that are relatively more isolated with less links to other regions.

$$NE = \begin{pmatrix} 1.533 & 0.869 & 1.148 & 1.406 & 1.456 & 1.451 & 1.456 & 1.533 \\ 0.869 & 1.419 & 0.804 & 1.303 & 1.404 & 1.410 & 1.417 & 1.494 \\ 0.996 & 0.767 & 1.414 & 0.855 & 1.310 & 1.388 & 1.411 & 1.489 \\ 1.398 & 1.289 & 0.847 & 1.437 & 0.868 & 1.302 & 1.397 & 1.480 \\ 1.480 & 1.397 & 1.302 & 0.868 & 1.437 & 0.847 & 1.289 & 1.398 \\ 1.489 & 1.411 & 1.388 & 1.310 & 0.855 & 1.414 & 0.767 & 0.996 \\ 1.494 & 1.417 & 1.410 & 1.404 & 1.303 & 0.804 & 1.419 & 0.869 \\ 1.533 & 1.456 & 1.451 & 1.456 & 1.406 & 1.148 & 0.869 & 1.533 \end{pmatrix} \quad (83.18)$$

The cumulative OE and DE matrices for the spatial model are shown in Eqs. (83.19) and (83.20), where we also see values that differ over the regions

and exhibit magnitudes greater than the 0.5 and 1.0 values representing coefficients β_o , β_d from the nonspatial model. As in the other cases, these reflect changes in flows arising from interactions modeled by origin-, destination-, and origin-destination dependence in the spatial gravity model. Intuitively, changing the characteristics of a single region will impact inflows to and outflows from that region, but also impact flows to regions neighboring the origin and impact flows to regions neighboring destination regions, and impact flows from regions neighboring the origin to regions neighboring the destination.

$$OE = \begin{pmatrix} 0.000 & 1.243 & 0.954 & 0.893 & 0.880 & 0.878 & 0.877 & 0.877 \\ 1.358 & 0.000 & 1.298 & 0.995 & 0.932 & 0.919 & 0.916 & 0.916 \\ 1.376 & 1.376 & 0.000 & 1.300 & 0.997 & 0.934 & 0.921 & 0.921 \\ 1.005 & 1.005 & 1.292 & 0.000 & 1.289 & 0.989 & 0.929 & 0.929 \\ 0.929 & 0.929 & 0.989 & 1.289 & 0.000 & 1.292 & 1.005 & 1.005 \\ 0.921 & 0.921 & 0.934 & 0.997 & 1.300 & 0.000 & 1.376 & 1.376 \\ 0.916 & 0.916 & 0.919 & 0.932 & 0.995 & 1.298 & 0.000 & 1.358 \\ 0.877 & 0.877 & 0.878 & 0.880 & 0.893 & 0.954 & 1.243 & 0.000 \end{pmatrix} \quad (83.19)$$

$$DE = \begin{pmatrix} 0.000 & 2.053 & 2.064 & 1.867 & 1.829 & 1.837 & 1.832 & 1.755 \\ 1.938 & 0.000 & 2.064 & 1.867 & 1.829 & 1.837 & 1.832 & 1.755 \\ 1.793 & 2.022 & 0.000 & 2.010 & 1.859 & 1.843 & 1.833 & 1.755 \\ 1.763 & 1.871 & 2.026 & 0.000 & 2.009 & 1.875 & 1.840 & 1.756 \\ 1.756 & 1.840 & 1.875 & 2.009 & 0.000 & 2.026 & 1.871 & 1.763 \\ 1.755 & 1.833 & 1.843 & 1.859 & 2.010 & 0.000 & 2.022 & 1.793 \\ 1.755 & 1.832 & 1.837 & 1.829 & 1.867 & 2.064 & 0.000 & 1.938 \\ 1.755 & 1.832 & 1.837 & 1.829 & 1.867 & 2.064 & 2.053 & 0.000 \end{pmatrix} \quad (83.20)$$

Using the same approach set forth in the previous section to produce scalar summary measures of the total effects, as well as the decomposition into origin, destination, intraregional, and network effects by averaging the matrices produced the results given in the second column of [Table 83.2](#).

The third and fourth columns show results based on calculating flow matrix responses to changes in each element of the $n^2 \times 1$ vectors X_o , X_d , which were added to produce the fifth column. In this case where a single characteristics vector x was used to form X_o and X_d , these equal the scalar summary effects produced by considering only n changes in elements of x_i .

Consider again the example involving commuting-to-work flows, where the number of residents is used as a size measure for origin regions and the number of business establishments as a size measure for the destination regions, so X_o and X_d are distinct. Interpreting results for this type of model would require reporting both columns three and four from [Table 83.2](#). Summing these two different scalar summary measures would make less sense in this situation, since changes in X_o do not imply changes in X_d and vice versa. This would lead to a slight change in interpretation, where changes in X_o (residents at the origin) lead to an origin, destination, intraregional, network, and total effects on flows, as do changes in X_d

Table 83.2 Scalar summary measures of effects for the spatial interaction model arising from a change in a single characteristic x averaged over all regions $i = 1, \dots, 8$

	Δx_i	$\Delta X_{o,i}$	$\Delta X_{d,i}$	$\Delta X_{o,i} + \Delta X_{d,i}$
Origin effects	0.9129	0.7920	0.1209	0.9129
Destination effects	1.6445	0.0605	1.5840	1.6445
Intraregional effects	0.3394	0.1131	0.2263	0.3394
Network effects	1.2698	0.4233	0.8466	1.2698
Total effects	4.1667	1.3889	2.7778	4.1667

(business establishments at the destination). This type of model specification could be viewed as an a priori zero restriction on the coefficient for the characteristic residents at the destination as well as a zero restriction on the coefficient for business establishments at the origin. It should be possible to include the full set of explanatory variables (residents and business establishments) in the set of model characteristics for both origins and destinations and then test the validity of the a priori zero restrictions. This would involve a test for significant differences between the full and nested model scalar summary effects estimates. If there are no differences in conclusions regarding the size and significance of the scalar summaries, then the restrictions are consistent with the sample data.

83.6 Conclusion

Recently introduced spatial autoregressive extensions of the spatial interaction model hold a great deal of promise for regional modeling of flows. However, there are still a great many obstacles to the wide use of these models in applied situations. First, these models require flow magnitudes that can be transformed to reflect a normal distribution. This is not the case for flow matrices containing a large number of zero values, large diagonal elements reflecting intraregional flows, or count magnitudes. There is a need for future research regarding implementation of a spatial autoregressive Poisson interaction model.

Beyond the issue of estimating model parameters, there is also a need to carefully consider how these parameters are interpreted. In the case of the independent spatial interaction model, changes in characteristics of a single region can exert impacts on inflows from all other regions, outflows to all other regions, as well as intraregional flows. These impacts can be measured by considering rows and columns of the flow matrix. We set forth a proposal for calculating scalar measures of impact that average over changes applied to a single explanatory variable (regional characteristic) for all regions. The approach allows separation of row/column and diagonal element impacts arising in the flow matrix, which we label origin, destination, and intraregional effects. Past applications of regression-based spatial interaction models that assume flows are spatially independent seem to have overlooked this aspect of the partial derivative impacts associated with changes in characteristics of regions.

For the case of the spatial autoregressive interaction model, interpretation of the model estimates in terms of their partial derivative impacts on flows is more complicated.

Changes to a single region's characteristics can impact not only inflows from all other regions, outflows to all other regions, and intraregional flows, but also all other flows in the flow matrix. These impacts can be measured using changes taking place in rows, columns, and the diagonal and off-diagonal elements of the flow matrix as a result of a change to a single region's characteristic. We propose a scheme for calculating scalar summary measures for these impacts that we label origin, destination, intraregional, and network effects.

Specifics regarding simulation of the partial derivative impact estimates based on the estimated distribution for the model parameters were not discussed here. This would require using the estimated variance-covariance matrix for the model parameters to generate draws for each model parameter. These could be used in conjunction with the approach proposed here to produce a distribution of the scalar estimates for the various types of impacts. These empirically derived distributions could serve as the basis for inference regarding significance of the various types of impacts.

References

- Behrens K, Ertur C, Koch W (2012) "Dual" gravity: using spatial econometrics to control for multilateral resistance. *J Appl Econom* 27(5):773–794. doi:10.1002/jae.1231
- Bolduc D, Laferriere R, Santarossa G (1992) Spatial autoregressive error components in travel flow models. *Reg Sci Urban Econ* 22(3):371–385
- Curry L (1972) A spatial analysis of gravity flows. *Reg Stud* 6(2):131–147
- Elhorst JP (2010) Applied spatial econometrics: raising the bar. *Spatial Econ Anal* 5(1):9–28
- Fischer MM (2002) Learning in neural spatial interaction models: a statistical perspective. *J Geogr Syst* 4(3):287–299
- Fischer MM, Griffith DA (2008) Modeling spatial autocorrelation in spatial interaction data: an application to patent citation data in the European Union. *J Reg Sci* 48(5):969989
- Fischer MM, Reismann M (2002) A methodology for neural spatial interaction modeling. *Geogr Anal* 34(2):207–228
- Fischer MM, Scherngell T, Jansenberger E (2006) The geography of knowledge spillovers between high-technology firms in Europe evidence from a spatial interaction modelling perspective. *Geogr Anal* 38(3):288–309
- Getis A (1991) Spatial interaction and spatial autocorrelation: a cross-product approach. *Environ Plan A* 23(9):1269–1277
- Gourieroux C, Monfort A, Trognon A (1984) Pseudo maximum likelihood methods: applications to Poisson models. *Econometrica* 52(3):701–720
- Griffith D, Jones K (1980) Explorations into the relationships between spatial structure and spatial interaction. *Environ Plan A* 12(2):187–201
- Lambert DM, Brown JP, Florax RJGM (2010) A two-step estimator for a spatial lag model of counts: theory, small sample performance and an application. *Reg Sci Urban Econ* 40(4):241–252
- Lee M, Pace RK (2005) Spatial distribution of retail sales. *J Real Estate Finance Econ* 31(1):53–69

- LeSage JP, Fischer MM (2010) Spatial econometric modeling of origin-destination flows. In: Fischer MM, Getis A (eds) *Handbook of applied spatial analysis*. Springer, Berlin/Heidelberg/New York, pp 409–433
- LeSage JP, Llano C (2006) A spatial interaction model with spatially structured origin and destination effects. SSRN: <http://ssrn.com/abstract=924603> or doi:10.2139/ssrn.924603. Accessed 17 Aug 2006
- LeSage JP, Fischer MM, Scherngell T (2007) Knowledge spillovers across Europe, evidence from a poisson spatial interaction model with spatial effects. *Pap Reg Sci* 86(3):93–421
- LeSage JP, Pace RK (2008) Spatial econometric modeling of origin-destination flows. *J Reg Sci* 48(5):941–967
- LeSage JP, Pace RK (2009) *Introduction to spatial econometrics*. Taylor-Francis/CRC Press, Boca Raton
- Porojan A (2001) Trade flows and spatial effects: the gravity model revisited. *Open Econ Rev* 12(3):265–280
- Ranjan R, Tobias JL (2007) Bayesian inference for the gravity model. *J Appl Econom* 22(4):817–838
- Roy JR, Thill JC (2004) Spatial interaction modeling. *Pap Reg Sci* 83(1):339–361
- Sen A, Smith TE (1995) *Gravity models of spatial interaction behavior*. Springer, Heidelberg
- Smith TE (1975) A choice theory of spatial interaction. *Reg Sci Urban Econ* 5(2):137–176
- Tiefelsdorf M (2003) Misspecifications in interaction model distance decay relations: a spatial structure effect. *J Geogr Syst* 5(1):25–50
- Wilson AG (1967) A statistical theory of spatial distribution models. *Transp Res* 1(3):253–269

Spatial models *price effect capital*
firms demand distribution modeling
knowledge growth *regression behavior*
location *goods innovation cities local*
models *cost distance information*
data *policy social environmental transport*
equilibrium
regions *travel markets area*
labor *population housing Space*
market *production impact wage migration supply trade*
network *countries geography*

Author Index

A

- Abdelmajid, M., 1360
Abdullah, S., 976
Abellan, J.J., 1279, 1369
Abraham, J.E., 743, 747
Abrahamsson, T., 762, 778
Abrahart, R.J., 1108, 1124
Abreu, M., 181, 304
Acemoglu, D., 171, 194, 195, 208, 853
Achour-Fischer, D., 129
Ackerman, F., 1051
Acs, Z.J., 206, 207, 209, 221, 381, 478, 481,
 500, 501
Acton, R.M., 1271
Adam, E., 1356
Adamic, L.A., 721, 827
Adamowciz, W., 986, 987
Adams, J.D., 408
Adams, S.A., 1327
Adnan, M., 1118, 1125
Aghion, P., 196, 197, 205, 215, 262, 333, 339,
 392, 534
Agrawala, S., 1059
Aigner, W., 1152
Aijun, D., 1020
Aitchison, J., 1466
Aitkin, M., 1340, 1354, 1500
Akaike, H., 1439
Akbari, H., 1052
Akella, M.R., 1386, 1394
Alan Glennona, J., 1313–1314
Alberini, A., 985
Albers, H.J., 1032, 1033, 1037, 1044, 1045
Albert, R., 729, 1268
Albrecht, G., 1545
Albrecht, J., 67
Alderson, D., 729
Aldstad, J., 1328
Alhassan, A., 1022
Alix-Garcia, J., 1041
Allen, C.D., 1206
Allen, P.M., 751
Allers, M.A., 1536, 1640
Alonso, W., 76, 514, 745, 749, 862
Alperovich, G., 696
Alvanides, S., 1168, 1169
Amin, A., 262, 269, 270, 272–274, 497
Amrhein, C., 1167
Anas, A., 552, 744, 745, 748, 817, 818, 838
Ancell, S., 1078, 1079
Andam, K., 1040, 1041
Anderson, B.,
 Anderson, D., 1340, 1354
Andersson, Å.E., 320, 431
Andersson, C., 1230
Andersson, F., 82
Andersson, M., 394, 399, 405, 409
Andienko, G., 712, 713
Ando, A.W., 1032, 1033, 1037, 1044, 1045
Ando, T., 1214
Andresen, M.A., 920
Andrews, F.M., 282
Andrienko, G., 1150, 1186, 1190, 1299
Andrienko, N., 712, 713, 1150, 1299
Angelescu, L., 280, 281, 284, 285
Angrist, J., 672, 673
Angulo, A., 1531
Ankerst, M., 1142
Anscombe, F.J., 1202, 1205, 1296
Anselin, L., 209, 500, 501, 979, 1031, 1035,
 1039, 1043, 1116, 1150, 1152, 1153, 1176,
 1283, 1285, 1286, 1289, 1305–1307, 1377,
 1379, 1421, 1512, 1515, 1521, 1525–1531,
 1536, 1537, 1540, 1562, 1567, 1572, 1588,
 1598, 1627, 1641, 1643, 1649
Apparicio, P., 1360
Appleyard, D., 1076
Aral, S., 721
Arbia, G., 1159, 1167, 1367, 1512, 1598, 1640
Archibugi, D., 384, 385

- Arellano, M., 305, 1644
 Arentze, T., 719, 720
 Arenze, T., 718, 722
 Armstrong, H., 265
 Armstrong, M., 1465, 1469
 Armsworth, P.R., 1044
 Arnold, R., 1448
 Arnott, R.J., 126, 552
 Aronsson, T., 936
 Arraiz, I., 1612
 Arrow, K.J., 417, 442, 443, 976
 Arthur, B., 610
 Arthur, M., 26
 Arthur, W.B., 610–615, 620
 Artis, M., 281
 Arundel, A., 384
 Aschauer, D., 340, 358
 Asheim, B.T., 377–379, 381–383, 385–387,
 452, 458, 462, 657
 Aslund, O., 101, 102
 Asplund, K., 1343
 Assuncao, R., 1425
 Atkinson, P.M., 1464, 1469–1471, 1625
 Audretsch, D.B., 202, 203, 206, 209, 214, 221,
 381, 422, 424, 425, 480, 482
 Auffhammer, M., 1039
 Aufhauser, E., 139, 151
 Autant-Bernard, C., 495, 660, 661, 1572
 AvRuskin, G., 1324, 1332
 Axelrod, A., 814
 Axelrod, R., 1227
 Axhausen, K., 720
 Axtell, R., 1227
 Aydalot, P., 445, 446
 Ayres, L.W., 1013–1016, 1018, 1019
 Ayres, R.U., 1013–1016, 1018, 1019, 1022,
 1025, 1026
 Azariadis, C., 307
- B**
 Backlund, K., 936
 Backlund, P., 1054
 Bacolod, M., 324, 325
 Badinger, H., 1612
 Bahn, O., 936, 938
 Bähr, C., 357, 358
 Bain, R., 700, 701
 Bairoch, P., 543, 551
 Baland, J.M., 1098
 Baldwin, R.E., 201, 216–219, 222, 226, 232,
 234, 542, 582, 586
 Balikcioglu, M., 942, 945
 Ballas, D., 283, 1238, 1240
 Baltagi, B.H., 1612, 1638, 1641–1643, 1646,
 1648, 1649
 Banergee, A., 674
 Banerjee, S., 1286, 1291, 1328, 1408, 1411,
 1447, 1572
 Banister, D., 747
 Banzhaf, H.S., 1042, 1043
 Bao, S., 52, 1306
 Barabási, A.-L., 721, 729, 826, 1268
 Barber, M.J., 482, 823
 Barca, F., 122, 535, 537, 663
 Barenklaau, K.A., 1037
 Bar-Gera, H., 761, 762, 769, 779–781
 Barkley, D.L., 52
 Barnard, S., 1238, 1246
 Barnes, S.A., 117
 Barnett, G.A., 737
 Barrat, A., 1271
 Barrett, S., 952, 954
 Barro, R.J., 171, 177, 180, 182–184, 189, 194,
 195, 262, 292, 298, 300, 304, 321, 864
 Barry, M., 1178, 1180
 Barry, R.P., 1566, 1574
 Bartelsman, E., 641
 Barthélémy, M., 738, 813, 827, 1265, 1271
 Basker, E., 70
 Bateman, I., 1035, 1038
 Bates, J., 692
 Batey, P.W.J., 894
 Bathelt, H., 492, 496, 592, 594–597, 599,
 600, 605
 Batta, R., 1386, 1394
 Batten, D.F., 406, 410, 888, 889
 Batty, J., 1114
 Batty, M.J., 816, 817, 821, 830, 837, 838, 849,
 1115, 1128, 1135, 1219, 1220, 1222–1224
 Batty, S.E., 1079
 Batz, M., 1044, 1045
 Baum, A., 127, 133, 134
 Baumann, J.H., 51, 1163
 Baumol, W.J., 195, 209, 483
 Baumont, C., 305
 Baum-Snow, N., 645, 646, 677
 Bayer, P., 1042
 Baylis, K., 1041, 1045
 Beardsley, K., 1420
 Beaudry, C., 485, 655
 Becattini, G., 443, 444
 Beck, N., 1420
 Becker, G.S., 118, 198, 262, 321, 327, 1096
 Becker, S.O., 357
 Beckman, R.J., 1236

- Beckmann, M.J., 238, 402, 511, 760, 761, 766, 779, 784, 789, 792, 793, 795, 807, 836
Begg, D., 122
Behar, J.V., 1391, 1393
Behrens, K., 567, 871, 1655, 1661
Behrens, W. III., 1089
Bekhor, S., 720
Belich, J., 337, 349
Bell, A.M., 285
Bell, B.S., 1301, 1424
Bell, D., 320
Bell, K.P., 1044, 1546–1548
Bell, M.G.H., 761
Belsley, D.A., 1444, 1445
Ben-Akiva, M., 709, 711, 717, 718
Benchekroun, H., 954, 960, 963
Bendtsen, P.R., 693
Benenson, I., 1224, 1227, 1229, 1230
Benko, G., 445
Bennett, J., 987, 990
Bennett, N., 1340
Bento, A.M., 1039
Bera, A., 1421, 1526, 1528
Bera, A.K., 1512, 1521, 1530
Berberoglu, S., 1464
Berchtold, B., 1142
Berger, G., 1004
Berger, M.C., 19, 26, 31, 160
Berger, T., 1227
Bergstrom, J., 989
Berk, R.A., 1439
Berke, P.R., 1072, 1073, 1079–1081
Berkman, L.F., 1336, 1341, 1344, 1345, 1351, 1352
Bernard, A., 307
Bernardin, V.L. Jr., 784
Bernardinelli, A., 1430
Bernardinelli, L., 1429
Beron, K.J., 1556, 1627, 1629
Berry, B.J.L., 284
Berry, C.R., 322
Berry, D., 1196
Bertin, J., 1145
Besag, J., 1213, 1281, 1409, 1410, 1413, 1415, 1424, 1505
Best, N.G., 1214, 1358, 1369, 1408, 1420, 1448
Beugelsdijk, M., 357
Bhargava, A., 1644–1645
Bhat, C.R., 707, 718, 1628, 1630, 1631
Bhattacharya, J., 262
Biehl, K., 1159–1161
Biggeri, A., 1431
Bimrose, J., 117
- Birch, K., 597, 612
Birdsall, N., 1088, 1089, 1091
Birkin, M.H., 1129, 1239–1241, 1245, 1251
Bishop, K.C., 1036
Bithell, M., 1230
Bivand, R.S., 1307, 1410, 1442, 1545, 1567
Black, D.A., 54, 996, 1375
Black, W., 1260
Blackburn, M.C., 1017
Blair, A., 1449
Blair, P.D., 887
Blakely, T.A., 1344, 1345
Blanchard, O.J., 68, 116
Blasius, B., 1271
Blatt, A., 1386, 1394
Blau, D.M., 64
Block, R.L., 134
Bloomquist, G.C., 19, 26, 31, 160, 1004
Blum, B., 324, 325
Blumenberg, E., 102
Blumsack, S., 1033
Blundell, R., 1644
Blyn, L.B., 1327
Blyth, B., 1238, 1246
Boccaletti, S., 823–825, 827
Bockstael, N.E., 998, 1044, 1546–1548
Boekema, F., 462
Boelhouwer, P.J., 141
Boggs, J.S., 592
Bogue, D.J., 7, 8, 10
Boianovsky, M., 170
Bolduc, D., 1655
Bollinger, C.R., 1612
Bolton, R., 1367
Bommaraju, T.V., 1017
Bonabeau, E., 719, 720
Bonanno, A., 47
Bond, S., 305, 1644
Bonini, A.N., 281, 282
Bonnin, E., 1473
Boon, F., 751
Boone, T., 792
Boots, B.N., 1265, 1268, 1482
Borcard, D., 1481
Borjas, G.J., 13, 853, 857
Borts, G.H., 18, 20, 23, 910
Boschma, R.A., 379, 381–383, 385–387, 467, 494, 498–500, 533, 535, 597, 599, 610, 626, 656–662
Boschma, R.M., 479, 485
Bosker, M., 532, 585, 1376
Boucher, A., 1114
Boustan, L.P., 12

- Bouwmeester, M.C., 888
 Bowden, R.J., 162
 Bowman, J.L., 709, 711, 717, 718
 Box, G., 1176, 1178
 Boxall, P., 1035, 1038
 Boyce, D.E., 761, 762, 769, 779–781, 784, 790,
 793, 888, 889
 Boyd, D.M., 732, 735
 Boyle, K., 985, 988
 Boyle, P., 1169
 Braczyk, H., 458, 461
 Bradford, S.C., 909
 Bradley, S., 114, 119
 Braess, D., 793, 799
 Brakman, S., 569, 571, 572, 578, 581–585
 Brander, J.A., 246
 Brandt, L., 646
 Brasier, K., 1100, 1101
 Brasington, D.M., 157, 158, 161, 162
 Brasington, J., 1230
 Braumoeller, B.F., 1341
 Braunerhjelm, P., 206, 381
 Brayne, C., 1358
 Breedon, D., 345
 Breschi, S., 418, 421, 422, 478, 479, 481, 658,
 660, 662
 Breshears, D.D., 1206
 Bresnahan, T., 596
 Brett, C., 1612
 Brewer, C.A., 1146, 1147, 1299
 Brewer, D., 721
 Brewer, I., 1189
 Brezger, A., 1409
 Briggs, D.J., 1279
 Bringezu, S., 1013
 Broemeling, L., 1424
 Brooks, K.R., 52
 Brookshire, D., 987
 Brown, D.F., 1078
 Brown, J.N., 157
 Brown, J.P., 1661
 Brown, L., 1236
 Brown, T., 987, 988
 Browne, W.J., 1347, 1353
 Brozović, N., 942
 Brueckner, J.K., 149, 1639
 Brülhart, M., 910, 918
 Brunsdon, C., 1187, 1188, 1289, 1394, 1436,
 1441, 1523
 Brusco, S., 453
 Bryk, A.S., 1345, 1354
 Buchanan, J.M., 420
 Buchert, M., 1020, 1021
 Buckeridge, D., 1313
 Bucovetsky, S., 968
 Buhl, S.L., 699
 Buliung, R.N., 719, 720
 Bullen, N., 1349, 1350
 Burdett, K., 63, 64
 Burger, H., 1466
 Burgess, T.M., 1392, 1393
 Burnett, K.M., 1033
 Burnett, W.S., 1320
 Burridge, P., 1562
 Busso, M., 676, 677
 Bussoletti, S., 357, 358
 Butcher, K.F., 14
 Button, K.J., 272, 273, 699, 839
 Butts, C.T., 1271
- C**
 Cahuc, P., 37, 40, 44, 45
 Cai, G., 1189
 Cairncross, F., 508, 726
 Caldarelli, G., 824, 825, 1268
 Calder, C.A., 1292, 1442, 1448
 Calthorpe, P., 1074
 Camagni, R., 387, 446, 492, 517
 Camann, D., 1449, 1450
 Cameletti, M., 1412
 Cameron, A.H., 1290
 Cameron, G., 140–142
 Cameron, T., 988, 996
 Campbell, A., 281, 284
 Campbell, M., 1086, 1087
 Campbell, S., 1073
 Can, A., 1522
 Cannari, L., 140, 142
 Cantril, H., 282, 288
 Canty, A., 1431
 Cao, Y., 1125
 Capello, R., 403, 495, 498, 519
 Cappelen, A., 357
 Carabelli, A., 444, 445, 452
 Caraça, J., 381
 Card, D., 14
 Carley, K., 1227
 Carlin, B.P., 1214, 1286, 1291, 1329, 1408,
 1411, 1447, 1572
 Carlin, J.B., 1408, 1411, 1576, 1580–1582,
 1584
 Carlino, G.A., 638, 639
 Carlo, M., 1225
 Carlsson, B., 206, 381
 Caroli, E., 205

- Carr, D.A., 1301
Carr, D.B., 1299–1302
Carree, M., 308
Carroll, C., 1420
Carroll, R.J., 1409
Carson, R.T., 976, 984, 987, 1089, 1092
Casado-Díaz, J.M., 49
Case, K.E., 140, 141
Casella, G., 1420
Casetti, E., 1289, 1522
Cass, D., 185
Castaldi, C., 610, 626
Castellacci, F., 357
Castells, M., 481, 497, 1254
Casti, J., 813, 814
Castles, F.G., 139
Cavaillès, J., 552, 554
Caves, R.E., 326
Ceccato, V., 1281, 1291
Ceh, B., 339
Cerhan, J.R., 1449, 1450
Chambers, R.G., 426
Champ, P., 987, 988
Chapin, F.S., 707, 745
Charlot, S., 635
Charlton, M.E., 1289, 1394, 1436, 1441, 1523
Château, J., 1057, 1060
Chatman, D., 1551
Chatterjee, S., 638, 639
Chavez, M., 823–825, 827
Chegwidden, J., 1020, 1021
Chen, G., 1020
Chen, J., 1557
Chen, J.T., 1336
Chen, J.X., 1301
Chen, Y., 563
Chenery, H.B., 889, 890
Cheng, T., 1177–1180, 1183, 1184, 1188, 1189, 1192
Cherry, T., 1100
Cheshire, P.C., 114, 158, 285, 670, 675
Chib, S., 1572, 1579, 1583
Chile, J.-P., 1462
Chimfwembe, K., 1420
Chiswick, B.R., 18
Chiu, S.N., 1265, 1268
Chopin, N., 1403–1406, 1412, 1415, 1432
Chorley, R.J., 727, 1255, 1265
Christakis, N.A., 721, 730, 731
Christakos, G., 1394, 1397
Chun, Y., 1481, 1489, 1500
Chunming, X., 1020
Ciccone, A., 476, 633, 634
Cingano, F., 641
Clark, A.E., 282, 285, 286, 1426
Clark, C.W., 933, 938, 941
Clark, L.C., 1320
Clark, M.J., 1625
Clark, T.N., 326
Clark, T.S., 1292
Clark, W.A.V., 150, 151, 279, 282, 285–287
Clarke, G.P., 1236, 1238, 1240, 1248
Clarke, H.R., 936
Clarke, K.C., 1114., 1220, 1223, 1224, 1369
Clarke, M., 1240, 1241
Clauß W., 1022
Clayton, D., 1429
Clayton, J., 129, 131
Cleveland, W.S., 1139, 1142, 1436
Cliff, A.D., 1165, 1167, 1281–1283, 1479, 1480, 1524, 1598, 1599, 1606
Clifford, N.J., 1230
Coale, A.J., 1089
Cochran, W.G., 1386
Coe, D.T., 409
Coe, N., 612
Coelho, D., 1052, 1057
Cohen, J.E., 1053
Cohen, M.D., 602, 814
Cohen, W.M., 379, 399, 417
Coleman, A., 335, 341
Collins, J., 1074
Collins, W., 12
Colt, J.S., 1449, 1450
Colwell, P.F., 129, 131, 132
Combes, P.-P., 476, 483, 486, 549, 551, 585, 587, 634, 636, 637, 642, 668
Comtois, C., 1271
Congdon, P.D., 1238, 1246, 1250, 1442, 1449
Conley, T.G., 1520, 1599, 1615
Conrad, J.M., 940, 942
Contractor, N., 721
Cook, D., 1305, 1306
Cook, K.A., 1189
Cooke, P., 379, 381–383, 385–387, 458, 461, 466, 470, 473, 490, 491, 657
Cope, M., 1119
Copeland, G., 1324, 1332
Corcoran, J., 1187, 1188
Cordy, C., 1494
Corfee-Morlot, J., 1057, 1060
Cornelius, I., 1022
Cornes, R., 416
Corsten, L.C.A., 1391
Cosmus, T., 1022
Costa, D., 87

- Costello, C., 941
 Couclelis, H., 716, 1224
 Coulson, N.E., 119
 Courchene, T.J., 116, 120
 Court, A.T., 994
 Courtat, T., 1270
 Cowan, S., 1073
 Cox, J.C., 65
 Cozen, W., 1449, 1450
 Cramer, J.S., 1557
 Crane, R., 1551
 Crecente, R., 1223
 Creel, M., 982
 Crescenzi, R., 650, 653, 659–661
 Cressie, N.A.C., 1198, 1208, 1279, 1284, 1292,
 1297, 1306, 1366, 1369, 1387, 1389, 1390,
 1394, 1424, 1464, 1598
 Crevoisier, O., 380–383, 445, 446
 Cronon, W., 1079
 Crooks, A., 1251
 Cropper, M.L., 934–936, 1034, 1035, 1037
 Crôte, A., 696
 Cruz, S.C.S., 447
 Cui, R., 1020
 Cukrowski, J., 907
 Cumbers, A., 597, 612
 Curran, P.J., 1464, 1471
 Currie, J.E., 1040
 Curry, L., 1654
 Cuzick, J., 1329, 1331
- D**
 Dachis, B., 677
 Dafermos, S.C., 238, 242, 761, 790, 792, 793,
 796, 802, 805, 806
 Dahl, G.B., 26
 Daily, G.C., 1044
 Dall’erba, S., 354, 357, 358, 360, 361, 363,
 364, 366, 367
 Dalmasso, P., 1431
 Dalton, R., 1387
 Dangelico, R.M., 499
 Daniels, T.L., 1073
 Dankbaar, B., 387
 Danuser, J., 1403
 d’Arge, R., 987
 Das, A.C., 1387
 Das, D., 1611
 Dasgupta, P., 839, 933, 934, 938, 1096
 d’Aspremont, C., 839, 847
 DaVanzo, J., 9
 David, P.A., 610–615, 617, 624
- Davidson, R., 1557, 1563
 Davies, L., 1243
 Davies, P.S., 862
 Davies-Withers, S., 150
 Davis, D.R., 576, 583, 584, 909
 Davis, S., 1021, 1449, 1450
 Davison, R., 938
 De Cea, J., 762, 780
 de Groot, H.L.F., 181, 203, 304, 357, 485,
 532, 655
 de Hoogh, C., 1279
 de Jong, P., 1480, 1482
 de Jong, R., 1587, 1588, 1612, 1644
 de la Barra, T., 747
 De la Roca, J., 637, 638
 De Maeyer, P.H., 720, 1230
 de Mooij, R., 357, 358, 361
 De Roos, A.J., 1449
 de Sherbinin, A., 1097, 1098
 de Smith, M.J., 1127, 1131, 1132
 De Souza Briggs, X., 1076
 de Zeeuw, A.J., 937, 939, 942, 963
 Deadman, P., 1115, 1227
 Deal, B., 1369
 Dearth, S., 1313
 Debarsy, N., 1429, 1591, 1643, 1645, 1649
 Decreuse, B., 89
 Dekock, V., 762, 780
 Delfiner, P., 1462
 Delmelle, E.M., 1386, 1388, 1391, 1393–1395,
 1397
 Delobel, R., 1020
 DeMaeyer, P., 715
 Deng, W., 781
 Denison, E.F., 195
 Denni, M., 384, 385
 Derudder, B., 1254
 Desbiens, C., 445
 Deshmukh, S.D., 936, 939
 Deskins, J., 1536
 Desmet, K., 234
 D’Este, P., 660, 661
 Deurloo, M.C., 279
 Deutsch, C.V., 1466, 1470
 Deutsch, S.J., 1177
 Dev, B., 312
 DeVany, A.S., 696
 Devlin, S.J., 1436
 Di Tella, R., 645
 Dial, R.B., 780
 Diamond, J., 1092
 Diamond, P.A., 60, 65, 68, 987
 Dieleman, F.M., 152, 279

- Dietz, T., 1094
Dietzel, C.K., 1220, 1369
Dietzenbacher, E., 899
Diggle, P.J., 1209, 1211, 1411–1413, 1420,
 1427, 1428, 1431
DiNardo, J., 676
DiPasquale, D., 129–131
Dittrich, S., 1020, 1021
Ditty, J., 1313
Dixit, A.K., 214, 215, 343, 540, 571, 912
Dixon, R., 264, 265
Dixon, T., 143
Dobler, C., 720
Dobrzynski, L., 1322
Dobson, A., 1343
Dockins, C., 995
Dockner, E.J., 946, 959, 962, 966, 967
Dodds, P.S., 730
Dodson, R.F., 1306
Doh, S., 699
Dolan, P., 288
Doloreux, D., 537
Domenich, T.A., 695
Domingo, R., 754
Don, C., 1077
Donaghy, K., 260
Donaldson, D., 646
Dong, J., 246, 790, 791, 794, 802, 805
Doppelhofer, G., 304
Döring, T., 653, 654
Dorling, D., 1238, 1240
Dornbusch, R., 122
Dosi, G., 394, 610, 626
Douady, S., 1270
Dowling, M., 494
Downs, A., 834
Doyle, J.C., 729
Doyle, J.J., 9
Dray, S., 1626
Drazen, A., 307
Dreassi, E., 1430
Drewe, P., 48
Drucker, P.F., 320, 495
Drukker, D.M., 1545, 1611–1614
Duan, W., 1021
Dubroca, L., 1473
Duckham, M., 1108
Duczmal, L., 1329
Duesenberry, J.S., 285
Duflo, E., 342, 674
Dugundji, E., 1626
Dujardin, C., 101, 103
Duncan, B., 1164
Duncan, C., 1344, 1345, 1357
Duncan, G.J., 1356
Duncan, O.D., 1164
Dungan, J.L., 1470
Dunk, J., 1420
Dupuis, P., 246
Durand, R., 627
Duranton, G., 87, 479, 486, 533, 634–637,
 642, 643, 645, 655, 656, 668, 671, 677,
 678, 1367
Durbán, M., 1409
Durbec, J.-P., 1473
Durlauf, S.N., 285, 286, 307
Dutt, A.K., 270
Dykes, J., 1187
- E**
- Earle, C.C., 1322
Easterlin, R.A., 6, 280, 281, 284–286, 1096
Eaton, B.C., 839
Eaton, J., 221
Echenique, M.H., 747
Ecker, D.J., 1327
Ecker, M., 1420
Eddington, P., 337
Ederveen, S., 357, 358, 361
Eding, G.J., 894
Edquist, C., 490, 656, 657
Edwards, M.E., 47
Edwards, R., 1093, 1329, 1331
Egenhofer, M.J., 1108
Egger, P., 308, 357, 1611–1614
Ehrlich, P.R., 1086
Ehud, O., 1227
Eichholtz, P., 129, 131
Eidsvik, J., 1403, 1411
Eijffinger, S., 357
Einstein, A., 818
Ekeland, I., 158
Elhorst, J.P., 1178, 1180, 1512, 1536, 1543,
 1640, 1643–1645, 1647, 1668
Ellegård, K., 707, 708, 720
Elliott, P., 1279, 1420
Ellis, J., 731
Ellison, G., 203, 533, 636
Ellison, N.B., 732, 735
Ellwood, J., 94
Elwood, S., 1119
Emmonds, A., 1188, 1192
Enayati, A., 1321
Engel, C., 547
Engelen, G., 834, 838, 1223

- Engelhardt, G.V., 140
 Engels, E.A., 1449
 Englund, P., 142
 Epple, D., 160
 Epstein, J., 1227
 Erdös, P., 825, 1265
 Erez, H., 1227
 Eriksson, A., 470
 Eriksson, R., 656
 Erlander, S., 770
 Erlich, A., 1089
 Erlich, P., 1089
 Ernst, D., 387
 Ertur, C., 181, 302, 305, 307, 366, 1429, 1572,
 1591, 1640, 1643, 1645, 1649, 1655, 1661
 Eshoo, M.W., 1327
 Esposti, R., 357, 358
 Essletzbichler, J., 51
 Esteban, A., 922
 Estevao, L.R., 1020
 Etilé, F., 282
 Evans, A.W., 128
 Evans, M.R., 1174, 1175
 Evans, S.P., 762, 768, 779, 780
 Evans, T.P., 1115
 Ewing, R., 1077
- F**
 Fackler, P.L., 942, 945
 Fagerberg, J., 357, 384, 385
 Faggian, A., 153, 221, 495, 498
 Fahrmeir, L., 1409
 Farber, S., 1443, 1444
 Farr, D., 1075
 Faust, K., 727, 728, 734, 1263
 Fayolle, J., 357, 358
 Feenstra, R.C., 571
 Fekete, J.D., 1190
 Feldman, M.P., 203, 214, 221, 382, 422, 424,
 425, 480, 482, 638, 642
 Ferdous, N., 1631
 Fernandez, C.E., 304
 Fernandez, J.E., 762, 780
 Ferraro, P.J., 1040, 1041
 Ferreira, F., 151
 Ferrer-i-Carbonell, A., 280
 Feser, E.J., 322, 324
 Few, S., 1150
 Field, E., 645
 Fieldhouse, E., 102
 Fields, G.S., 120
 Figueiredo, O., 533, 640
 Filer, R.K., 13
 Filippetti, A., 384, 385
 Filzmoser, P., 1466
 Findeis, J., 1100, 1101
 Finelli, L., 1313
 Fingleton, B., 267, 302, 303, 306, 313, 531,
 573, 585, 1286, 1537, 1611
 Finley, A.O., 1443, 1448
 Fiorello, D., 755
 Fischer, M.M., 51, 139, 151, 152, 161, 163,
 181, 190, 302–304, 312, 321, 365, 381, 405,
 482, 493–495, 501, 531, 585, 821, 823, 907,
 1035, 1124, 1129, 1150, 1163, 1181, 1197,
 1255, 1279, 1285, 1298, 1359, 1360, 1375,
 1500, 1572, 1654, 1655, 1661, 1662
 Fischer, S., 122
 Fishback, P.V., 12
 Fisher, A., 941
 Fisher, J.D., 129
 Fisher, L.M., 119
 Fitjar, R.D., 596
 Fitoussi, J.P., 278
 Flanagan, K., 387
 Flatau, P., 89
 Flatman, G.T., 1391, 1393
 Florax, R.G.J.M., 1526, 1528, 1529
 Florax, R.J.G.M., 181, 304, 1529, 1661
 Flores, N., 984, 987
 Florian, M., 761, 780
 Florida, R., 203, 209, 319, 320, 322–327, 339,
 462, 476, 482, 536
 Flowerdew, R., 1162, 1165–1167, 1500
 Flyvbjerg, B., 699, 700
 Foi, A., 1202
 Folmer, H., 1529
 Forbes, M., 89
 Forman, C., 339
 Fornahl, D., 470, 598
 Fornalski, K.W., 1322
 Forrest, J., 1163
 Forslid, R., 201, 216–219, 222, 226, 234, 542,
 574, 586, 668
 Fortunato, S., 1263
 Fossett, M., 285, 286
 Foster, J., 616
 Foster, N., 909
 Fotheringham, A.S., 816, 1116, 1162, 1163,
 1165, 1166, 1289, 1306, 1394, 1436, 1441,
 1442, 1523, 1625
 Fowler, J.H., 721, 730, 731
 Frank, R.H., 285
 Franzosa, R.D., 1108
 Fratesi, U., 357

- Freedman, M., 637
Freeman, M., 976
Freeman, R.B., 383
Frenchman, D., 712
Frenken, K., 467, 482, 484, 486, 494, 498, 535,
 593, 594, 597, 626, 660–663, 1230
Frentzos, E., 713
Frey, B.S., 279
Frey, H., 1077
Frey, W.H., 13
Friedman, J.W., 239
Friedman, M., 698, 700
Friesz, T.L., 238, 239, 243, 244, 246,
 250, 251
Frijters, P., 280, 285, 286
Fritsch, M., 481
Fu, X., 693
Fuentes, M., 1433
Fuglstad, G.A., 1412
Fujita, K., 1017
Fujita, M., 222, 223, 337, 356, 515, 528,
 531, 541, 542, 559, 576, 583, 643,
 839, 1367
Fulton, W., 1074
Funk, S., 1323
Furnas, G.W., 1149
Fürst, F., 743, 744
- G**
Gabe, T., 324
Gabszewicz, J., 839, 847
Gahegan, M., 1189
Gaigné, C., 552, 554, 563, 564, 566
Galdo, J., 996
Gallagher, C.M., 1329
Gallego-Fernandez, J., 1403
Galliani, S., 645
Gallup, J.L., 1094
Galor, O., 1091
Gambardella, A., 596
Gamerman, D., 1575–1578
Gandolfo, G., 819, 821
Gandy, A., 1556
Ganeshan, R., 792
Gannon-Rowley, T., 1341
Garavelli, A.C., 499
Garcia, A.M., 1223
Garcia-Milà, T., 359
Garcia-Penalosa, C., 205
Gardner, M., 1221
Garlick, J., 1387
Garofalo, G.A., 361
- Garretsen, J.H., 532, 569, 571, 572, 578,
 581–585, 615
Garrett, T.A., 360, 363, 364, 366
Garud, R., 619, 620, 624
Gaspar, J., 476
Gassler, H., 51
Gatrell, A.C., 1386
Gatzlaff, D.H., 159
Gaudet, G., 963
Gauthier, D., 120
Gautier, P.A., 86, 87
Gayda, S., 754
Gaydos, L., 1114, 1223
Gayer, T., 1040
Gazel, R., 920
Gazulis, N., 1220, 1369
Gebreab, S., 1427
Geddes, A., 1166
Geddes, P., 1079
Geels, F., 466
Geenhuizen, M.S., 495
Gehlke, C.E., 1159–1161
Geisser, S., 1407
Gelfand, A.E., 1292, 1408, 1411, 1420, 1447,
 1572, 1576
Gell-Mann, M., 1219
Gelman, A., 283, 1402, 1403, 1576,
 1580–1582, 1584
Geltner, D.M., 129, 131
Geman, D., 1575
Geman, S., 1575
Gennetian, L., 1356
Gentleman, R., 1307
Geoghegan, J., 1034
George, A., 1566
George, E., 1420
Gerking, S., 51
German, S.E., 1625
Geroski, P.A., 394
Gertler, M.S., 377, 378, 381, 458, 594
Getis, A., 161, 163, 1035, 1116, 1150, 1165,
 1314, 1359, 1360, 1478, 1481, 1626, 1654
Geweke, J., 1556, 1574, 1585
Geyer, C., 1575, 1582, 1583
Ghislandi, M., 1429
Ghosh, A., 876, 898–900
Giannotti, F., 716
Giarratani, F., 47
Gibbons, S., 672–675
Giddens, A., 594, 597, 601, 604
Giesen, C., 200
Gilbert, N., 1192, 1227, 1238
Gill, H.L., 161

- Gillis, C.R., 1164
Gillman, R., 984
Gimblett, H.R., 1227
Gin, A., 84
Girvan, M., 1263
Gitelman, A., 1427, 1428
Giuliani, E., 492, 493
Gjerde, J., 936
Glaeser, E.H., 121, 203, 334, 337
Glaeser, E.L., 36, 84, 87, 320–322, 325, 326,
 422, 424, 442, 476, 478, 479, 485, 532, 533,
 543, 636, 637, 639, 640, 645, 857
Gleditsch, K., 1420
Gloaguen, C., 1270
Glückler, J., 592, 596, 597, 599, 605
Glymour, M.M., 1344, 1345
Gnad, F., 750
Gobillon, L., 95, 98, 103, 585, 634, 636, 637,
 642, 678
Godin, B., 652
Goeschl, T., 941, 945
Goetz, S.J., 36, 48, 52, 1100
Goldberg, M.A., 134
Goldfarb, A., 339
Goldstein, H., 1345, 1347–1349, 1352, 1353
Goldstein, N.C., 1220, 1369
Gollier, C., 940
Gómez-Rubio V., 1410
Goodchild, M.C., 1290, 1366, 1369
Goodchild, M.F., 1108, 1110, 1115, 1116,
 1119, 1120, 1127, 1131, 1132, 1313–1314
Goodrich, C., 5
Goovaerts, P., 1281, 1291, 1328, 1329,
 1393–1395, 1397, 1427, 1462, 1464,
 1469–1471, 1473, 1474
Gopinath, M., 341
Gordon, I., 69, 70, 675
Gordon, P., 1550
Görg, C., 1190
Gorman, D., 1443, 1444
Gorter, C., 82
Gorter, J., 357, 358, 361
Goss, E., 120
Goto, A., 377
Gottlieb, J.D., 121, 334, 532, 533, 645
Gotway, C.A., 1328, 1443, 1444, 1470
Gourieroux, C., 1657
Goyal, S., 736
Grabher, G., 593, 596, 600, 603, 604, 610
Gradus, R., 262
Graedel, T.E., 1024
Grafton, R.Q., 1032, 1045
Graham, C., 288
Graham, D.J., 485, 696
Graham-Tomasi, T., 940
Gramlich, E., 333
Granovetter, M., 593, 594, 728
Gråsjö, U., 405
Graves, P.E., 18, 23, 24, 26, 996, 999,
 1001, 1005
Gray, W., 1034
Green, A., 117
Green, M., 1165, 1166
Greenberg, E., 1579
Greenberg, M., 1076
Greenberger, M., 752
Greene, S.K., 1133
Greene, W.H., 1621, 1623, 1624
Greenhut, M., 238
Greenland, S., 1329
Greenstein, S., 339
Greenstone, M., 633
Greenwood, M.J., 5, 8, 11, 13, 14, 23,
 852, 862
Gregory, J., 676, 677
Greiling, D., 1464
Grepperud, S., 936
Griese, H., 1010
Griffin, W.A., 1115
Griffith, D.A., 1177, 1178, 1386, 1443, 1444,
 1478, 1480–1482, 1489, 1494, 1496, 1500,
 1562, 1567, 1626, 1655
Griliches, Z., 376, 377, 407, 415, 418, 419, 994
Grimes, A., 335, 340, 342
Gronau, R., 63
Gross, J.L., 1255
Gross, T., 1271
Grossman, G.M., 196, 197, 202, 215, 216,
 218–221, 234, 423, 426, 1092
Grubel, H.G., 571, 910
Gruenewald, P., 1443, 1444
Guerrieri, P., 444
Guimaraes, P., 533, 640
Guinet, C., 1473
Guiyuan, J., 1020
Gulden, T., 339
Guo, D., 914, 918
Guo, J., 914
Gupta, C.K., 1022–1023
Gurak, D.T., 14
Guthrie, G., 343, 345
Gutmann, M., 721
Gutwin, C., 1149
Guy, F., 660, 661
Gwilliam, K.M., 696
Gyourko, J., 334, 337

H

- Ha, C., 1548, 1549
Haab, T., 982, 983, 985, 988
Hadjimichalis, C., 594
Haenlein, M., 731–733
Hageluken, M., 1010
Hagen, T., 357, 358, 363, 364
Hägerstrand, T., 402, 405, 690, 707, 708, 714,
 738, 745, 1153, 1188, 1369
Haggett, P., 727, 1255, 1265
Haining, R.P., 1280, 1281, 1284, 1286–1288,
 1290–1292, 1386, 1393, 1462, 1481, 1537,
 1562, 1598
Hajivassiliou, V., 1556
Hakimi, S.L., 244
Hakley, M., 1230
Hall, P., 1074
Hall, R.E., 633, 634
Hall, T.A., 1327
Hallegatte, S., 1057, 1060
Halsema, A., 963
Haltiwanger, J., 82, 641
Hamilton, B.W., 78, 79, 86
Hammond, G.W., 1376
Han, J., 1174, 1190, 1290
Han, Y., 1100
Handbury, J., 548, 554
Handfield, R.B., 789
Hanemann, M., 941, 976, 984, 987
Hannis, J.C., 1327
Hansen, D.L., 736
Hansen, W.G., 743
Hanson, S., 1057, 1060
Hanushek, A., 1040
Hanushek, E., 151, 152
Hao, Y., 1329
Hardin, C., 1146
Harding, A., 1236
Harding, D., 101
Hardisty, F., 1189
Hardy, M.H., 699
Harker, P.T., 238, 240, 244
Harland, K., 1129, 1236, 1240, 1241, 1245,
 1246, 1248, 1250
Harmaakorpi, V., 386, 460, 463
Harris, B., 751
Harris J.R., 855, 856
Harrower, M.A., 1146, 1147, 1299
Hartge, P., 1449, 1450
Harvey, J., 128
Hashimoto, H., 244
Haslett, J., 1307
Hasnain-Wynia, R., 1314
Hassink, R., 592, 594, 626
Hastings, W.K., 1575, 1580
Hatna, E., 1230
Hatton, T.J., 18
Haughwout, A., 339
Haurie, A., 936, 938
Haurin, D.R., 159
Hausman, J., 987
Hausner, J., 274
Haustein, J., 1022
Haworth, J., 1178, 1179
Hällervik, A., 1230
He, S., 780
Head, K., 576–577, 582, 583
Heal, G.M., 930, 933–935, 938, 944, 945
Hearn, D., 761
Heasman, M.A., 1164
Hecht, J., 1057
Heckman, J.J., 158
Hedlund, G.A., 1221
Heffley, D., 836
Heggie, I., 691
Heidenreich, M., 458, 460, 461
Heijdra, B.J., 571
Held, L., 1403, 1406, 1407, 1410, 1413
Helliwell, J.F., 283
Helpman, E., 196, 197, 202, 215, 216,
 218–221, 234, 409, 418, 423, 426, 542
Hemingway, J., 1321
Hendershott, P.F., 159
Hendershott, P.H., 89
Henderson, D., 310
Henderson, J.M., 238, 239
Henderson, J.V., 429, 633, 645, 646
Henderson, R., 415, 419, 429
Henderson, V., 204, 396, 404, 543, 1375
Hengl, T., 1393
Henley, A., 140, 142
Henn, S., 470
Henning, J.A., 994
Henry, M.S., 52
Henry, S., 1099
Hensher, D., 986, 988
Heppenstall, A.J., 1129, 1240, 1245, 1251
Heras, H.E., 936
Hermkens, K., 732–734
Herriges, J., 983
Herweijer, C., 1057, 1060
Heuvelink, G.B.M., 1177, 1178, 1181
Hewings, G.J.D., 914, 915, 918–920
Hey, A.J.G., 1109
Heylighen, F., 813
Hicks, J.H., 544

- Hicks, J.R., 4, 10
 Higdon, D., 1428
 Higgs, G., 1187, 1188
 High, J., 64
 Hill, B., 1536
 Hillberry, R., 922
 Hinde, J., 1340, 1354
 Hirsch, G., 444, 445, 452
 Hirst, P., 274
 Hite, D., 157, 161, 162
 Hoeffler, A., 305
 Hoehn, J.P., 19, 26, 31, 160, 990, 1004
 Hoermann, S.A., 7, 8, 10
 Hoeting, J., 1427, 1428
 Hoffmann, M.J., 1115, 1227
 Hofman, F., 718, 720
 Hofstadler, S.A., 1327
 Höhle, M., 1329
 Holdren, J.P., 1086
 Hole, D.J., 1164
 Holland, J.K., 1219, 1220
 Holland, M., 1206
 Hollanders, H., 384
 Holm, M., 699
 Holmes, M.J., 140, 141, 1081
 Holmes, T., 986, 987
 Holt, J., 1165
 Holtz-Eakin, D., 359–361
 Holzer, H., 102
 Hondronyiannis, G., 1540
 Honey-Roses, J., 1041, 1045
 Hong How, H.H., 134
 Hoover, E.M., 47, 834, 840, 1089
 Hoover, K.D., 170
 Hoozemans, F.M.J., 1052
 HöpfI, H., 469
 Hopkins, R.S., 1313
 Hoppen, S., 1223
 Horan, P.M., 48
 Hornbeck, R., 633
 Hossain, M., 1420
 Hotelling, H., 567, 835–837, 839, 932
 Howe, H.L., 1320
 Howitt, P., 196, 197, 215, 262, 333, 339,
 392, 534
 Hoyler, M., 1254
 Hron, K., 1466
 Hsiao, C., 1644
 Hsieh, C.-T., 641, 642
 Hsieh, W.W., 1182, 1183
 Huang, B., 1178, 1180
 Huang, H.-J., 780
 Huang, J.-C., 988, 1254
 Huang, Q., 1125
 Hudak, S., 1306
 Hudson, G., 1462
 Hughes, G., 118
 Hugo, G., 1092, 1093, 1099
 Huijbregts, C.J., 1462, 1465, 1469
 Huismans, G., 755
 Huizinga, H., 969
 Hummel, J., 1313
 Hummels, D., 913, 918, 922
 Hung, C.S., 238
 Hung, N.M., 945
 Hunt, G.L., 13, 23
 Hunt, J.D., 743, 747
 Hunt, R.M., 638, 639
 Hunter, L.M., 1093, 1095, 1100, 1102
 Hürlimann, E., 1420
 Hussain, A., 1227
 Huybers, P., 1433
 Hwang, D.U., 823–825, 827
 Hyndman, R.J., 1201
- I**
- Iammarino, S., 444, 447, 535, 597–599, 657,
 660, 661
 Ihaka, R., 1307
 Ihlanfeldt, K., 95, 98, 99, 103, 105
 Iida, Y., 761
 Illian, J.B., 1403, 1413
 Inglehart, R., 282, 327
 Inoguchi, Y., 1021
 Inselberg, A., 1142
 Ioannides, Y.M., 142
 Irvine, K., 1427, 1428
 Irwin, E.G., 1044, 1368
 Isaaks, E.H., 1469
 Isard, W., 508, 511, 791, 815, 816, 818,
 882, 890
 Ishii, R., 643
 Islam, N., 304
 Ismaila, A., 1431
 Israilevich, P.R., 914
 Isserman, A., 51
 Iwano, E.J., 1320
- J**
- Jackman, R., 70, 71
 Jackson, R.W., 890
 Jackson, S.K., 1227
 Jacobs, A., 1076

- Jacobs, J., 318, 319, 323, 325, 326, 403, 404, 422, 424, 476, 484, 485, 1075, 1076
Jacobson, M., 1322
Jacquez, G.M., 1324, 1329, 1332, 1464
Jaffe, A.B., 408, 415, 418, 419, 429
Janelle, D.G., 1120
Janikas, M.V., 1175
Jansen, A., 1323
Jansenberger, E., 1661
Janssen, M.A., 1115, 1227
Jarup, L., 1279
Jayaraman, V., 792
Jayet, H., 1643, 1649
Jeannerat, H., 380–383
Jebara, T., 721
Jemal, A., 1329
Jenish, N., 1512, 1615
Jenkins, G., 1176, 1178
Jennish, N., 1538, 1545
Jennrich, R., 1557
Jensen-Butler, C., 896
Jenson, R., 969
Jeong, H., 1268
Jerry, W., 1077
Jiang, B., 1270
Joh, C.-H., 718, 722
Johansson, B., 394, 399, 400, 406, 408–410, 415, 419, 426
Johansson, C., 1308
Johnson, D., 1420
Johnson, G.E., 112, 1313
Johnson, K.P., 49
Johnson, P., 300, 307
Johnson, S., 195, 208
Johnston, R., 1163
Johnston, S.C., 1323
Jones, C.I., 1589, 1590
Jones, C.R., 1449
Jones, J.P. III., 1289
Jones, K., 1336, 1344, 1345, 1349, 1350, 1353, 1357, 1655
Jones, P.M., 692, 707
Jones, R.W., 913, 923
Jonhson, P., 307
Jons, H., 482
Jonsson, O., 379
Joppa, L., 1040
Jorgensen, S., 946, 952, 959, 962, 970
Journel, A.G., 1181, 1462, 1465, 1466, 1469, 1470, 1473
Jovanovic, M., 540
Jowsey, E., 128
Juarrero, A., 472
Judge, G.G., 789
Jul, S., 1149
Jungnickel, D., 1255
Juniper, J., 270, 272, 273
- K**
Kabatereine, N., 1420
Kaddour, A., 1345
Kahn, J., 985
Kahn, M.E., 84, 87, 1040
Kahneman, D., 279
Kain, J., 94, 97, 98, 101, 699, 701
Kain, J.F., 82
Kaiser, B.A., 1033
Kakamu, K.W., 1429, 1572
Kaldor, N., 195, 215, 265, 428
Kallal, H.D., 203, 442, 485, 532, 640
Kamarianakis, Y., 1177, 1180
Kamien, M.I., 938
Kammann, E.E., 1408, 1409
Kan, K., 150
Kanaroglou, P.S., 719, 720
Kanevski, M., 1182, 1183
Kang, J.M., 1174, 1175
Kansky, K., 1255, 1257, 1259
Kantor, Y., 89
Kaplan, A.M., 731–733
Kapoor, M., 1612
Kareiva, P., 1044
Karetnikov, D., 1052, 1057
Karl, T.R., 1058
Karlsson, C., 418, 424
Karlsson, S., 394, 399
Karnøe, P., 619, 620, 624
Karp, L., 937, 941
Kartha, S., 1051
Katherine Yih, W., 1329
Katz, L.F., 104, 116, 1356
Kauffman, S., 458, 468, 469
Kawachi, I., 1336, 1341, 1344, 1345, 1351, 1352
Keane, M., 1556
Keeble, D., 446, 451
Keilbach, M., 381
Keim, D.A., 1142, 1190
Keith, B., 1077
Kelejian, H.H., 1519, 1522, 1525, 1529, 1530, 1540, 1598, 1599, 1603, 1607, 1608, 1610–1612, 1641
Keller, W., 409
Kelley, H., 1115
Kemeny, J., 139

- Kemp, M.C., 934
 Kemp, W., 1164
 Kendal, M.G., 1159
 Kennedy, P., 989
 Kennell, D.L.,
 Kent, P., 134
 Kenworthy, J.R., 1076, 1077
 Kerr, W.R., 533, 636, 639, 640, 643
 Kerry, R., 1281, 1291, 1462
 Kessler, R.C., 1356
 Ketels, C., 454
 Khandker, S.J., 1038
 Kho, Y., 1153, 1307
 Kierzkowski, H., 913, 923
 Kietzmann, J.H., 732–734
 Kilkenny, M., 361
 Kim, C.W., 1540
 Kim, K.H., 142, 143
 Kinderlehrer, D., 804
 King, G., 721, 1114
 King, L.J., 726, 1388
 Kingsley, B.S., 1312, 1323
 Kingsnorth, D., 1020, 1021
 Kinkela, S.B., 1323
 Kirat, T., 499
 Kirby, A.M., 1160
 Kirby, D.K., 1549
 Kirkwood, G.P., 938, 941
 Kisalu, T.L., 1323
 Klaassen, L.H., 48, 1283, 1598
 Klaerding, C., 592, 594
 Klaiber, H.A., 1043
 Klaver, J., 1369
 Kleinman, K., 1329
 Klenow, P.J., 641, 642
 Klepper, S., 598
 Klette, T.J., 395
 Klier, T., 1628
 Kline, P., 676, 677
 Kling, C., 983
 Kling, J.R., 104, 1356
 Klomp, A., 1189
 Klomp, L., 308
 Knapp, K., 939
 Kneib, T., 1409
 Knorr-Held, L., 1406
 Kobayashi, T., 715
 Koch, J., 620, 624
 Koch, W., 181, 302, 366, 1572, 1640,
 1655, 1661
 Kockelman, K., 1572, 1627–1629, 1632
 Koethenbuerger, M., 969
 Kogler, D.F., 382
 Kohlhammer, J., 1190
 Kolko, J., 320, 325, 326, 479, 532, 543, 678
 Kompas, T., 1032, 1045
 Konduri, K., 1631
 Koolwal, G.B., 1038
 Kooperberg, C., 1424
 Koopmans, T.C., 185, 836
 Kopp, R., 976, 984
 Koppelman, F., 784
 Korbel, J., 752
 Kort, J.R., 49
 Kortum, S., 221, 395
 Koschinsky, J., 460, 461, 1380
 Kraak, M.J., 1189
 Kraft, K., 695
 Krieger, N., 1336, 1345
 Krige, D., 1473
 Krishnamurthy, N., 1022–1023
 Kristensen, T., 1420
 Kriström, B., 930, 944, 945
 Kritz, M.M., 14
 Kroes, E.P., 691
 Krueger, A.B., 279, 1092
 Krugman, P.R., 200, 201, 207, 215, 221–225,
 233, 262, 268, 273, 337, 356, 429, 476, 479,
 528, 530, 531, 540–542, 549–551, 559, 566,
 570, 571, 576–583, 585, 595, 610, 611, 617,
 645, 668, 815, 839, 869, 907, 911–913, 916,
 1093, 1366, 1367
 Krukow, K., 734
 Krumm, R., 999
 Krutilla, J., 976
 Kuhn, H.W., 764
 Kuijpers, B., 713–715
 Kulldorff, M., 1133, 1288, 1329
 Kumaraswamy, A., 619, 620, 624
 Kutner, M.H., 1443
 Kutzbach, M., 82
 Kuznets, S., 5
 Kverndokk, S., 936
 Kwan, M.-P., 715
 Kyriakidis, P.C., 1114, 1181, 1471, 1474
- L**
- LaCombe, D., 161, 163
 Laferriere, R., 1655
 Lagazio, C., 1430
 Lago, A.M., 697
 Laibman, D., 262
 Lajaunie, C., 1290
 Lakatos, I., 528
 Lakshmanan, T.R., 272, 273

- Lall, S.V., 360, 363, 364
Lam, W.H.K., 780
Lambert, D.M., 1661
Lambert, S.,
Lambooij, J., 485
Lamorgese, A.R., 567
Landström, H., 443, 444
Lane, W.J., 238
Lang, S., 1409
Lansing, J.B., 9
Larch, M., 1567
Larsen, K., 1341
Larson, D., 980
Latora, V., 738, 823–825, 827
Lautso, K., 754
Lawson, A.B., 1328, 1420, 1426, 1432
Lawton Smith, H., 381, 387
Layard, P.R.G., 88, 112
Lazer, D., 721
Le Bras, M., 1020
Le Gallo, J., 302, 303, 305, 354, 357, 358, 361,
 363, 364, 366, 367, 1368, 1372, 1375–1377,
 1643, 1649
Leamer, E.E., 496, 653, 907, 976
Lechner, C., 494
Léchot, G., 445, 451
Lecoq, B., 445, 451
Lecuyer, A., 357, 358
Lee, D.B., 747, 1550
Lee, D.-H., 781
Lee, D.J., 1409, 1425, 1426
Lee, D.S., 676, 677
Lee, K., 969
Lee, L.-F., 1587, 1588, 1599, 1609, 1611,
 1612, 1641, 1643, 1644, 1649
Lee, M., 1655
Legendre, P., 1481, 1626
Lehmann, E., 209
Leinberger, C., 1078
Leithead, D.J., 1567
Leithead, W.E., 1567
Leizarowitz, A., 942
Lemieux, T., 677
Lemp, J., 1572, 1627, 1632
Lengerich, E., 1189
Lenzi, C., 660, 662
Leonard, D., 961
Leontief, W.W., 875–877, 890, 909
Lerch, F., 597
LeRoy, S., 83
LeSage, J.P., 152, 161, 301, 303, 304, 308, 364,
 365, 482, 660, 661, 1116, 1191, 1291, 1360,
 1368, 1420, 1421, 1423, 1442, 1490, 1500,
 1514, 1516, 1520, 1523, 1525, 1537, 1538,
 1540, 1541, 1543, 1545, 1547–1549, 1551,
 1556, 1561, 1563, 1564, 1572, 1573, 1583,
 1586–1589, 1591–1593, 1598, 1612, 1625,
 1627–1629, 1632, 1641, 1643–1645, 1649,
 1654–1662, 1664, 1666, 1667
Lessard, D., 342
Leung, Y., 821, 1124, 1442, 1444
Levin, D., 1648, 1649
Levin, R.C., 399
Levine, R., 304
Levinson, D.A., 1271
Levinthal, D.A., 379, 417, 602
Lewin-Koh, N., 1305, 1306
Ley, E., 304
Leydersdorff, L., 495, 834
Li, D., 1646
Li, H., 54, 55, 862
Li, P.-F., 594, 600
Li, W., 1125
Li, X., 1115, 1177, 1180
Li, Z., 1020, 1184
Liang, Y., 340
Liaw, K.L., 13
Liebman, J.B., 104, 1356
Lin, J., 639, 641
Lin, X., 1609, 1612, 1641
Lindgren, F., 1411–1413, 1415
Lindgren, U., 656
Lindsey, R., 688
Lindström, J., 1411–1413, 1415
Ling, D.C., 159
Linneman, P.D., 18, 27
Lipsey, R., 839
Lissoni, F., 418, 421, 422, 478, 479, 481, 658
List, J.A., 967
Liu, J., 1566
Liu, L., 996, 1303
Liu, R., 1020, 1021
Liu, X., 1612, 1641
Liu, Y., 748, 838, 1432
Livingston, J.M., 1449
Llamosas-Rosas, I., 360, 363, 364, 366
Llano, C., 922, 1661
Lloyd, A., 1323
Lloyd, C.D., 1289, 1464, 1466, 1468
Lloyd, J.O., 1323
Lloyd, P.J., 571, 910
Lloyd, R., 1376
Lockwood, B., 969
Löfgren, K.G., 936
Long, F., 1443
Long, J.A., 713, 714

- Long, N.V., 934, 935, 946, 952, 954, 959–963,
966, 967, 970
- Longley, P.A., 738, 1110, 1114, 1116, 1118,
1127, 1131, 1132, 1219
- Lööf, H., 394, 399
- Loomis, J., 979, 980, 982, 984, 987, 989, 990
- Lopes, H.F., 1575–1578
- López, E., 755
- Lopez, F., 1531
- Lopez, R.A., 47
- Lopez-Bazo, E., 302
- Lorrain, F., 1264
- Lösch, A., 422
- Lott, N., 1057
- Louviere, J., 986, 988
- Loveridge, S., 359
- Lowe, S.E., 1039
- Lowry, I.S., 746
- Lu, H., 1286, 1291, 1329
- Lubart, T.I., 323
- Lubchenko, J., 1086
- Lubin, L., 1449, 1450
- Lucas, R. Jr., 198–200, 215, 319, 326, 426
- Lucas, R.E., 87, 88
- Ludwig, J., 1356
- Lüke, B., 1017
- Lundqvist, L., 762, 778
- Lundvall, B.-Å., 381, 426, 495
- Lung, Y., 499
- Lunn, D., 1448
- Luo, J., 1625
- Lutz, W., 1092
- Lynch, C.F., 1449
- Lynch, J., 1343
- Lynch, K., 1078
- Lytinen, S.L., 1227
- M**
- Ma, B., 1432
- MacEachren, A.M., 1138, 1145, 1189
- Macedo, J., 713
- Machlup, F., 320
- Machnes, Y., 696
- Mack, E., 1380
- MacKay, D., 1432
- MacKaye, B., 1079
- Mackie, P.J., 696
- MacKinnon, D., 597, 612
- MacKinnon, J., 1557, 1563
- MacLaren, A.M., 1164
- MacLennan, D., 136
- Macy, M., 721
- Madow, L.H., 1386
- Madow, W.G., 1386
- Madsen, B., 896
- Maffi, L., 1146
- Maggioni, M.A., 660, 661
- Magnac, T., 103
- Magnani, C., 1431
- Magrini, S., 670
- Maguire, D.J., 1110, 1115, 1116
- Maguire, K., 995
- Mahmassani, H.S., 793
- Maier, G., 663
- Maignan, C., 495
- Maillat, D., 445, 446, 451
- Maimbo, S.M., 872
- Mairesse, J., 495
- Majumdar, M., 938
- Majure, J.J., 1305, 1306
- Mäkitalo, M., 1202
- Makower, H., 4, 7, 8, 10, 11
- Malecki, E.J., 495, 653, 654, 1271
- Malerba, F., 627, 657
- Malhamé, R., 936, 938
- Malmberg, A., 492, 496, 594–598
- Mandalaz, D., 1199
- Manduchi, A., 418, 424
- Mangel, M., 930, 939
- Mankiw, N.G., 181, 292, 321
- Manley, D., 1167
- Manley, E., 1192
- Mann, J.R., 1290
- Mansfield, E., 396, 398, 406
- Manski, C.F., 675
- Manson, S.M., 1115, 1227
- Manta-Conroy, M., 1072
- Mantsinen, J., 431
- Marans, R.W., 284
- Marble, D., 1303
- Marchand, M., 1052
- Marcotte, P., 761, 784
- Mardia, K.V., 1426, 1562
- Marion, J., 677
- Markandya, A., 976, 1051
- Markoff, J., 729
- Markovich, S., 246
- Markusen, A., 657, 668
- Marmot, M., 285
- Marques, C., 755
- Marrocu, E., 660, 662, 663
- Marschak, J., 4, 7, 8, 10, 11
- Marshall, A., 53, 85, 429, 441–443, 451–453,
483, 491
- Marshall, E.C., 1407

- Marshall, R.J., 1562
Marshall, S., 747
Marston, A., 143
Marston, S.T., 111, 120
Martin, D., 1169
Martin, K., 984, 987
Martin, P.H., 215–217, 219, 222, 226, 232,
 233, 542, 574, 586, 668
Martin, R., 114, 116, 269, 270, 273, 379,
 381–383, 385–387, 448, 462, 466, 468, 470,
 485, 593, 594, 596–598, 604, 668
Martin, R.J., 1566, 1567
Martin, R.L., 610, 611, 614, 615, 621, 626
Martin, S.W., 1165
Martinez, F.J., 749
Martin-Herran, G., 952, 959, 970
Martino, A., 754
Martino, S., 1403–1406, 1410–1412, 1415,
 1432
Mascolo, C., 738
Maskell, P., 492, 496, 595–598
Maskin, E., 839
Mason, C.F., 380, 967
Masoumi, A.H., 791, 807
Massard, N., 495
Massey, D., 320
Massire, C., 1327
Matas, A., 102
Matasci, G., 1182
Matérn, B., 1391
Matheron, G., 1208, 1281, 1388
Matheronw, G., 1473
Matsumura, T., 238
Matsushima, N., 238
Maule, M., 1431
Maxwell, J., 1138
Mayer, T.H., 549, 551, 576–577, 582, 583
Mayo, S.K., 137
Mayworm, P., 697
McAllister, P., 143
McBratney, A.B., 1392, 1393, 1464, 1465,
 1473
McCall, B.P., 69
McCall, J.J., 61, 69
McCann, P., 122, 153, 184, 215, 221, 263,
 318, 319, 447, 485, 528, 530, 531, 535, 537,
 573, 598
McCarthy, I.P., 732–734
McCollum, D., 987, 988
McCombie, J.S.L., 267, 268
McConnell, K.E., 982, 983, 985, 998
McCormick, B., 118, 119
McDonald, J.F., 128, 1081, 1625
McDonnell, S., 1228
McDowell, M., 916
McEnroe, J.M., 697
McFadden, D., 690, 744, 815, 818, 1556, 1622
McGee, R., 114, 115, 117
McGranahan, D., 324
McGuire, C.B., 760, 766, 779, 789, 792, 793,
 795, 807
McGuire, T.J., 359
McKinnish, T., 54
McMahan, J., 127
McMaster, R., 597, 612
McMillen, D.P., 126, 128, 1528, 1625, 1628
McNally, M.G., 692, 707, 719
McNeill, J.R., 1086, 1095
Meadows, D., 1089
Meen, G., 137, 140
Megretskaia, I., 1305, 1306
Mei, C.L., 1442, 1444
Melançon, G., 1190
Meliker, J.R., 1324, 1329, 1332
Melillo, J.M., 1058, 1086
Melitz, M.J., 582, 907, 911, 912
Melkas, H., 460, 463
Mellander, C., 339
Mellinger, A.D., 1094
Melo, P.C., 485
Mendonça, S., 381
Menezes, R., 1412
Meng, Q., 781
Meng, X.L., 1584–1586
Menzel, M.-P., 470, 598
Merchant, C., 1073
Merchante, A.J., 281
Merletti, F., 1431
Merlo, J., 1341, 1343
Merz, C., 1020, 1021
Metcalf, D., 115
Metcalfe, J.S., 480, 616
Metropolis, N., 1575, 1577, 1579
Michel, P., 1165
Michelacci, C., 640
Middendorf, A., 1010
Middleton, D., 1165
Mieszkowski, P., 968, 969
Miksch, S., 1152
Mila i Canals, L., 1015
Milgram, S., 728–730
Miller, E.J., 752
Miller, H.J., 713–716, 1174, 1190, 1220, 1225,
 1290, 1381
Miller, N.G., 129, 131
Miller, R.E., 304, 342, 887

- Mills, J.A., 76, 1512, 1538
 Minard, C., 1138
 Mincer, J., 52, 321, 871
 Minshull, R., 1387
 Mirabelli, D., 1431
 Miranda, D., 1223
 Mitchell, R., 976, 984
 Mitchell, W., 270, 272, 273
 Mitleton-Kelly, E., 463, 465, 467, 469,
 472, 473
 Mitra, A., 543
 Miyamoto, K., 749, 1444
 Moffitt, R., 120
 Mohan, P., 1174, 1175
 Mohl, P., 357, 358, 363, 364
 Molho, I., 69
 Mollie, A., 1409, 1410, 1413, 1415, 1425
 Monastiriotis, V., 285
 Monestiez, P., 1473
 Monfort, A., 1657
 Monk, S., 138
 Monmonier, M.S., 1147, 1149, 1186, 1298,
 1299, 1303
 Mønsted, M., 494
 Montgomery, J., 1075
 Montomoli, C., 1429
 Monzon, A., 755
 Moodysson, J., 379, 381–383, 385–387
 Mookherjee, R., 246
 Moon, G., 1238, 1246, 1336, 1344, 1345, 1357
 Mooney, H.A., 1086
 Moore, S.A., 1079
 Moran, P., 1523, 1524
 Moreno, Y., 823–825, 827
 Morenoff, J.D., 1341
 Moretti, E., 52, 633
 Morgan, J.P., 699, 1020–1023
 Morgan, K., 462, 491
 Moriguchi, Y., 1013
 Moriset, B., 495
 Morris, D.R., 1024, 1026
 Morris, S., 1420
 Morrison, C., 340
 Morrison, D., 1058
 Morrison, P.S., 281, 287
 Mortensen, D., 60, 61, 63, 65
 Morton, L.M., 1449
 Moses, L.N., 889, 890
 Mostashari, F., 1329
 Motohashi, K., 377
 Mouzon, S.A., 1078
 Moyeed, R., 1209, 1428
 Muellbauer, J., 140–142
 Mueller, E., 9
 Mueller, J., 979, 1010
 Mueser, P.R., 23
 Muir, K.R., 1290
 Muir-Wood, R., 1057, 1060
 Mulembakani, P.M., 1323
 Müller, W., 1152, 1386
 Munch, J.R., 88, 89
 Munnell, A.H., 358–361, 363
 Munroe, D.K., 914, 918
 Mur, J., 1531
 Murdoch, J.C., 999
 Murmann, P., 465
 Murphy, A.D., 1036
 Murphy, F.H., 241
 Murray, A.T., 1379
 Murray, I., 1432
 Mushinge, G., 1420
 Musolesi, M., 738
 Muth, R.F., 13, 76, 514
 Mutl, J., 1612
 Myers, D.E., 1391
 Myrdal, G., 268, 271, 428, 541
 Mytelka, L.K., 378, 381, 385
- N**
- Nachtshheim, C.J., 1443
 Nagaoka, S., 377
 Nagurney, A., 238, 242, 243, 246, 761,
 788–791, 793, 794, 796, 799, 801–805,
 807, 808
 Nahuis, R., 357, 358, 361
 Nascimento, R.S.V., 1020
 Nash, J., 66, 239, 240, 243, 245, 246, 249,
 250, 257
 Nash, J.F., 796
 Nauwelaers, C., 386, 459
 Nævdal, E., 934
 Navratil, F.J., 9
 Neary, J.P., 587
 Neidell, M., 1040
 Neill, D.B., 1184, 1185
 Nelson, J., 989
 Nelson, R.R., 221, 392, 394, 417
 Nelson, T.A., 713, 714
 Neprash, J.A., 1282
 Nesheim, L., 158
 Neter, J., 1443
 Neumark, D., 678
 Neutens, T., 715, 720, 1230
 Neutra, R.R., 1313
 Newburn, D.A., 1044

- Newhouse, J.P., 1322
Newland, D., 453, 454
Newman, D.H., 942
Newman, M.E.J., 730, 1255, 1263, 1271
Newman, P.W.G., 1076, 1077
Nguyen-Luong, D., 753
Ni, P., 1254
Niazi, M., 1227
Nicholls, D.C., 134
Nicholls, R.J., 1052, 1057, 1060
Nichols, E.L. Jr., 789
Nicholson, W., 120
Nickel, E., 839
Nickell, S.J., 88
Nicolis, G., 822
Nie, Y., 781
Nielsen, M., 734
Nielsen, S.B., 969
Niendorf, H.P., 1022
Nijkamp, P., 89, 90, 184, 190, 299, 495, 813,
 816, 818–821, 823, 839, 1481
Nilsson, L., 918
Nizalov, D., 359
Nohria, N., 602
Noland, R.B., 485, 696
Nolles, K., 1033
Nonaka, I., 478
Nordbeck, S., 1260
Nordhaus, W.D., 944, 1090
Norman, G., 238
Norman, V., 571
North, C., 1147
North, D.C., 499
Nosvelli, M., 660, 661
Novshek, W., 238
Nriagu, J., 1324, 1332
Ntoutsi, I., 713
Nucci, F., 140, 142
Nuijten, A., 755
Nunes, P., 976
Nyström K., 398
- O**
Oakes, J.M., 1356
Oates, W.E., 158, 1034, 1035, 1037
Oaxaca, R.L., 65
O'Brien, T.F., 1017
Obstfeld, M., 907, 911, 912
Odoi, A., 1165
O'Donoghue, C., 1238
Ohlin, B., 403, 404, 791
O'hUallachain, B., 339
- Okabe, A., 1265, 1268
Okazaki, F., 914
O'Kelly, M.E., 816
Okubo, T., 582
Okulicz-Kozaryn, A., 283, 284
Okuyama, Y., 914, 919, 920
Olea, R.A., 1393, 1394, 1397
Oliver, M.A., 1213, 1281, 1290, 1291, 1462,
 1468, 1473
Olmstead, S., 1034
Olsen, A., 1386
Olson, L., 939
Oltvai, Z.N., 826
O'Malley, A.J., 1354
Oosterhaven, J., 882, 884, 887, 888, 890, 891,
 893, 894, 898, 900
Oosterlynck, S., 612
Openshaw, S., 1108, 1113, 1124, 1133,
 1158–1161, 1168, 1169, 1240, 1299
Oppenheim, N., 761, 778
Orcutt, G., 752
Ord, J.K., 1116, 1165, 1167, 1281–1283,
 1479, 1480, 1524, 1537, 1554, 1559,
 1560, 1562, 1565, 1568, 1598,
 1599, 1606
Orr, L.L., 158
Ortega, B., 281
Ortiz, R.A., 1051
Orton, M., 117
Ortúzar, J.D., 761, 769
Osborne, M.J., 66
Osth, J., 101, 102
O'Sullivan, D., 1224, 1230, 1265
Oswald, A.J., 88, 119, 279
Othman, W., 713–715
Ottaviano, G.I.P., 215–217, 219, 222,
 226, 232, 233, 495, 540, 542,
 548, 567, 574, 586
Ottaviano, G.M., 668
Otten, R.H.J.M., 1243
Oum, T.H., 693, 694
Oury, K., 1081
Overman, H.G., 116, 334, 336, 486, 533, 668,
 671–675, 678, 1367
Overton, W.S., 1387, 1388
Owen, D., 1075
Owen-Smith, J., 496, 595
Ozgen, C., 184
- P**
Pace, K.P., 301, 364, 365, 1420, 1421, 1423,
 1490, 1514, 1520

- Pace, R.K., 161, 308, 1116, 1191, 1360, 1500, 1512, 1516, 1523, 1525, 1537, 1538, 1540, 1541, 1543, 1545, 1547, 1548, 1551, 1556, 1561, 1563, 1564, 1566, 1567, 1572–1574, 1583, 1586, 1591, 1593, 1598, 1627–1629, 1632, 1641, 1645, 1654–1662, 1664, 1666
- Paci, R., 660, 662, 663
- Paelinck, J., 1283, 1367
- Paelink, J.H.P., 1598
- Páez, A., 1443, 1444
- Page, S., 618
- Palander, T., 425, 834, 840
- Paleti, R., 1631
- Palley, T., 270, 271
- Palmquist, R.B., 996, 1035, 1042
- Pande, R., 342
- Papacharissi, Z., 738
- Parent, O., 1512, 1538, 1572, 1587–1593, 1612, 1644, 1649
- Parizh, M., 1022
- Parker, D.C., 1115, 1227
- Parkes, S.E., 1290
- Parkinson, S., 140, 141
- Parmesan, C., 1054
- Parmeter, C., 310
- Parr, J.B., 915, 920
- Parsons, G., 982, 983
- Partridge, M.D., 22, 54, 55, 111, 115, 116, 119, 120, 154, 155, 163, 1044, 1368, 1372
- Pascutto, C., 1429
- Pastor-Satorras, R., 1271
- Patmore, N., 1057, 1060
- Patriksson, M., 761, 762, 766, 769, 773, 780, 784, 790
- Patuelli, R., 1481
- Paul, C., 120
- Paul, M., 1329
- Paul, S.,
- Pawlowsky, V., 1466
- Pearce, B., 138
- Pearce, F., 1057
- Pearce, J.R., 1236, 1240, 1241, 1245, 1246, 1250
- Pearson, K., 1139
- Peasgood, T., 288
- Pebesma, E.J., 1410
- Peca, S.P., 128, 129
- Pedreschi, D., 716
- Pehkonen, J., 112
- Pekkola, S., 386
- Pelekis, N., 712, 713
- Pendyala, R., 1631
- Peng, H., 1021, 1023
- Pentland, A., 721
- Percival, R.,
- Peres-Neto, P.R., 1626
- Perez, J., 922
- Perino, G., 941, 945
- Perkins, H.C., 148, 149
- Pesaran, M.H., 1644
- Peterson, T.C., 1058
- Petrongolo, B., 67, 68, 86
- Petruzzielli, A.M., 499
- Pettit, L.I., 1407
- Pfaff, A., 1040, 1041
- Pfaffermayr, M., 308, 1612
- Pfeifer, P.E., 1177–1179
- Pfister, M., 445, 451
- Phaneuf, D.J., 1043
- Phillips, D., 82
- Phinikettos, I., 1556
- Phipps, T.T., 1540
- Pickle, L.W., 1299, 1301
- Pickles, J., 1118
- Pickrell, D.H., 699, 700
- Pietrobelli, C., 444
- Pike, A., 269, 597, 612
- Pike, W., 1189
- Pillot, C., 1021
- Pimentel, D., 1058
- Pindyck, R.S., 343, 930, 938–940, 942, 945
- Pinelli, D., 495
- Pinjari, A.R., 707, 718
- Pinkse, J., 1612, 1628
- Piore, M.J., 318, 320, 443, 444
- Piras, G., 1529, 1640
- Pirotte, A., 305, 306
- Pischke, J.S., 672, 673
- Pissarides, C.A., 60, 65, 67, 86, 116
- Pittau, M.G., 283
- Platt, J.R., 1314, 1329
- Platteau, J.P., 1098
- Pliska, S.R., 936, 939
- Plourde, C., 939
- Polanyi, M., 402
- Polasek, W., 1572
- Polasky, S., 937, 1032
- Polèse, M., 537
- Pollakowski, H., 82
- Ponds, R., 482, 660–663
- Poot, J., 184, 190, 203, 299, 485, 532, 655
- Popp, D., 1090
- Porojan, A., 1655
- Porter, M.E., 318, 325, 447, 448, 450, 451, 453, 454, 492, 494, 534, 645
- Porter, R.H., 359

- Portney, P., 976, 987
Portnoy, S., 1495
Postel-Vinay, F., 282
Poulsen, M., 1163
Powell, W.W., 496, 595
Pozdnoukhov, A., 1182
Prastacos, P., 1177
Pred, A., 707, 720
Prescott Adams, R., 1432
Presser, S., 976, 984
Price, V.C., 916
Prigogine, I., 822
Proctor, J.D., 1116
Prucha, I.R., 1512, 1522, 1538, 1545, 1598,
 1599, 1608, 1610–1615, 1641
Puerto, J., 839
Puga, D., 53, 87, 116, 479, 533, 549, 581, 582,
 585, 633, 635, 637, 638, 642, 643, 645, 655,
 656, 870
Pulido, A., 922
Pulselli, R.M., 712
Purves, R.S., 1182
Putnam, R., 274, 535, 536
Puu, T., 238, 843
Pyatt, G., 891
Pyka, A., 502
Pyke, F., 453
- Q**
Qiang, Q., 788, 791, 799
Quah, D.T., 309, 1374, 1375
Quandt, R.E., 238, 239
Quarendon, G., 1238, 1246
Quigley, J., 140, 141, 151, 152, 400
Quyen, N.V., 945
- R**
Raa, T., 238
Rabe, B., 149
Rabellotti, R., 444, 445, 452
Rabier, J.R., 282
Raciborski, R., 1545
Radner, R., 976
Railsback, S.F., 1227
Rakodi, C., 1076
Ramirez, I., 1041, 1045
Ramlogan, R., 616
Ramos, P.N., 923
Ramsey, F., 185
Ran, B., 790
Randers, J., 1089
- Ranjan, R., 1661
Rantisi, N.M., 592
Rao, L., 1168
Raphael, S., 102
Rapoport, D., 913, 918
Rappaport, J., 154, 163
Rasbash, J., 1347, 1353
Rasmussen, E., 834, 844
Råstam, L., 1343
Ratcliffe, J., 129, 1421
Ratha, D., 872
Ratti, C., 712
Rauch, J., 322
Raudenbush, S.W., 1345, 1354
Ravenstein, E.G., 4, 6, 743
Raymond, J.-L., 102
Read, D., 1231
Reaser, J., 102
Recker, W.W., 718, 719
Redding, S.J., 531, 584, 585
Redmond, G.,
Reed, W.J., 729, 936–939, 941, 942, 1368
Rees, P.,
Reeson, A.F., 1033
Reggiani, A., 813, 816, 818–824, 829, 830
Rehkopf, D.H., 1336
Reich, B., 1433
Reichl, H., 1010
Reilly, C.S., 1286, 1291
Reimann, C., 1466
Reinganum, J.F., 962, 963
Reismann, M., 1654
Ren, C., 342
Renaud, B., 142, 143
Rényi, A., 825, 1265
Rephann, T.J.,
Resseger, M.R., 637, 640
Revelt, D., 304
Revilla Diez, J., 381
Rey, S.J., 302, 312, 313, 366, 367, 1175, 1368,
 1372, 1375–1377, 1379, 1380, 1529
Reynold, C., 1227
Reza, A.M., 697
Rhind, D.W., 1110
Ribeiro, P.J., 1211, 1411, 1413, 1420,
 1427, 1428
Ricardo, D., 512, 513
Rice, P., 334, 336
Richards, K., 1230
Richardson, H.W., 1550
Richardson, S., 1279, 1369, 1420
Rickman, D.S., 54, 55, 111, 115, 116, 119, 154,
 155, 163, 1100, 1368, 1372

- Ridker, R.G., 994
 Riebler, A., 1403
 Riefler, R., 890
 Rietveld, P., 726, 738
 Rigdon, M.A., 246
 Rimoin, A.W., 1323
 Rindt, C.R., 692, 707, 719
 Riou, S., 564, 566
 Ripley, B.D., 1278, 1296, 1389
 Risser, P.G., 1206
 Riva, M., 1360
 Rivkin, S.G., 1040
 Rivlin, A., 752
 Roback, J., 18–20, 22, 279, 326, 333, 994
 Robalino, J.A., 1040, 1041
 Robert-Nicoud, F., 542, 574, 586, 668, 871
 Robert-Nicoud, R., 216, 217, 219, 226
 Robinson, A., 1387
 Robinson, D.P., 1519, 1525, 1530
 Robinson, H.W., 4, 7, 8, 10, 11
 Robinson, J., 195, 208
 Robinson, P.M., 1615
 Rodriguez, J.P., 1271
 Rodrigues, A., 1425
 Rodriguez, L.C., 1033
 Rodriguez-Pose, A., 117, 122, 357, 535, 537,
 596, 650, 653, 659–661
 Rogers, E.M., 418
 Rogers, J., 547
 Rogerson, P.A., 69, 1328, 1386, 1394
 Roig, J.-L., 102
 Rolfe, J., 987, 990, 1033
 Romaya, S., 1076
 Romer, D., 181, 292, 321
 Romer, P.M., 196, 197, 200, 201, 206, 215,
 216, 218, 392, 426, 427, 442, 443
 Roos, M., 1407
 Root, G.S., 719
 Rosa, E.A., 1094
 Rose, A., 829
 Rosen, H.S., 157
 Rosen, S., 156–158, 160, 333, 994
 Rosenbaum, J.E., 1356
 Rosenberger, R., 989
 Rosenbloom, J.L., 18
 Rosenblueth, A., 1220, 1221
 Rosenbluth, A.W., 1575, 1577, 1579
 Rosenbluth, M.N., 1575, 1577, 1579
 Rosenthal, S.S., 53, 563
 Rosholm, M., 88, 89
 Rosling, H., 1308
 Rospabé, S., 103
 Ross, S., 97
 Ross, T., 1057
 Rossi, P.H., 149
 Rossi-Hansberg, E., 87, 88, 234
 Rossiter, D.G., 1238, 1240, 1393
 Rothman, N., 1449, 1450
 Rothschild, M., 64
 Round, J.I., 891
 Rouwendal, J., 81, 82, 90
 Roux, A.V.D., 1337, 1341
 Roux, S., 642
 Rovinsky, R.B., 244
 Rowthorn, R., 270
 Roy, D., 721
 Roy, G.G., 838–841
 Roy, J.R., 1654
 Royuela, V., 281
 Rubin, C.H., 1312, 1323
 Rubin, D.B., 1408, 1411, 1576,
 1580–1582, 1584
 Rubinstein, A., 66
 Rubio, S., 952, 954, 955, 957, 958
 Rue, H., 1403–1407, 1410–1413, 1415, 1432
 Rueda-Cantuche, J.M., 888, 890
 Rupasingha, A., 52
 Ruppert, D., 1409
 Russell, R., 310
 Russo, D., 1391
 Ruth, M., 1052, 1057
 Ruttan, V.W., 652
 Rutten, R., 462
 Ruud, P., 976, 984
 Ryznar, R., 1081
- S**
- Sabel, C.F., 318, 320, 443, 444, 461, 463
 Sacerdote, B., 674
 Sachs, J.D., 1094
 Sage, J., 1429
 Saiz, A., 320–322, 325, 326, 479, 532, 543
 Saks, R.E., 12, 334, 337
 Sala-i-Martin, X., 171, 177, 180, 182–184, 189,
 194, 195, 292, 298, 300, 304, 864
 Salais, R., 595, 601
 Salathe, M., 1323
 Salcedo Du Bois, R., 1101
 Salvini, P.A., 752
 Samand, H.A., 1038
 Sampath, R., 1327
 Sampson, R.J., 1341
 Samuelson, P.A., 540, 789
 Sanbonmatsu, L., 1356
 Sanchez-Azofeifa, A.G., 1040, 1041
 Sanchirico, J.N., 1044

- Sanders, M., 202, 206, 207
Sanders, S., 54
Sanderson, W.C., 1092
Sandler, T., 416
Sang, E., 103
Sanglier, M., 751
Santarossa, G., 1655
Sante, I., 1223
Saphores, J.D., 942
Sargan, J.D., 1644–1645
Sattenspiel, L., 1323
Savinov, A., 1299
Savouri, S., 70
Saxenian, A., 318, 493, 494, 534, 594, 596, 597
Scafidi, B., 98
Scarpetta, S., 641
Scellato, S., 738
Schabbenberger, O., 1470
Schargrodsky, E., 645
Scharnhorst, A., 502
Scheinman, J.A., 203, 442, 485, 532, 640
Schelling, T.C., 285, 286, 1227, 1229
Scherbov, S., 1092
Scherngell, T., 482, 823, 1654, 1661
Schiffauerova, A., 485, 655
Schiller, R., 1020
Schindler, G.R., 914
Schivardi, F., 641
Schleifer, A., 203, 442, 485
Schlenker, W., 942
Schmeichel, K.L., 1312, 1323
Schmidt, M.A., 1133
Schneider, V., 1032, 1045
Schnellenbach, J., 653, 654
Schödle, B., 1407
Schofield, D.,
Schönhofen, H., 139
Schoumaker, B., 1099
Schramm, M., 583–585
Schréyogg, G., 469, 620, 624
Schroedle, B., 1403
Schubert, U., 51, 1163
Schüler, D., 1020, 1021
Schulze, W., 987
Schumann, H., 976, 1152
Schumpeter, J.A., 392, 393, 399, 403, 478,
 480, 484
Schur, N., 1420
Schwanen, T., 720
Schwartz, A., 340
Schwartz, H.M., 139, 140, 143
Schwartz, N.L., 938
Schwarz, D., 379, 381–383, 385–387
Schweizer, U., 839
Scott, A.J., 325, 602
Scott, J.P., 727, 728, 734, 823
Seabrooke, L., 139, 140
Seabrooke, W., 134
Sear, D.A., 1625
Searle, B.A., 140, 141
Selod, H., 95, 98, 99, 101, 103
Seltén, R., 961
Semboloni, F., 834, 838
Sen, A., 278, 817, 818, 1654, 1656
Sener, I., 1631
Sengenberger, W., 453
Senik, C., 282
Sestito, P., 140, 142
Sethi, G., 941
Setterfield, M., 266, 268, 274, 612, 613, 615
Severson, R.K., 1449, 1450
Sexton, R.L., 26
Shapiro, J.M., 532
Shaw, R.P., 120
Shaw, S.-L., 716
Shearmur, R., 537, 1360
Sheffi, Y., 761, 773, 790
Sheils, J.F.,
Shekhar, S., 1174, 1175, 1190
Sheldon, R.J., 691
Shepherd, M., 129
Sheppard, I., 754
Sheppard, S., 153, 158, 285
Sherali, H.D., 241
Shields, M.A., 285, 286
Shiller, R., 140, 141
Shimpo, K., 909
Shioji, E., 359
Shleifer, A., 532, 640
Shneiderman, B., 736, 1147, 1149
Shoemaker, C.A., 244
Shogren, J.F., 1032, 1033, 1037, 1044, 1045
Shreve, R.L., 1114
Shryock, H.S. Jr., 7, 8, 10
Siderius, W., 1390, 1391
Sieg, H., 160
Siegfried, J., 337
Silva, E.A., 1223
Silva, O., 640
Silvestre, B.S., 732–734
Simmie, J., 446, 462
Simmonds, D.C., 749
Simon, H., 814
Simon, J., 1090
Simon, L., 840
Simon, N., 995

- Simoonga, C., 1420
 Simpson, D., 1412, 1413
 Singleton, A.D., 738, 1118
 Siow, A., 999
 Sizer, M., 49, 50
 Sjaastad, L.A., 18, 70, 71, 865
 Sjoquist, D., 95, 103, 105
 Skopik A., 1149
 Slack, B., 1271
 Slade, M.E., 1612, 1628
 Slavtchev, V., 481
 Slingsby, A., 1187
 Slobierski, M.G., 1133
 Slotnick, M.J., 1324, 1332
 Small, C., 1053
 Small, K.A., 79, 552
 Smart, M.W., 48
 Smirnov, O., 1567
 Smit, M.J., 203, 485, 532, 655
 Smith, A.F.M., 318, 325, 483, 1576
 Smith, D.M., 1129, 1236, 1240, 1241, 1245,
 1246, 1248, 1250
 Smith, J.S., 140, 141
 Smith, K., 378, 381, 385
 Smith, M.A., 736, 761
 Smith, M.J., 802
 Smith, N.K., 1323
 Smith, R.H.T., 726
 Smith, T.E., 817, 818, 1423, 1627, 1628, 1632,
 1654, 1656
 Smith, V.K., 1038, 1039
 Smulders, S., 262
 Snickars, F., 381, 744, 836, 838–841
 Snow, J., 143, 1138, 1298, 1299
 Soete, L., 381
 Solow, R.M., 170, 174–177, 195, 262, 355,
 366, 976
 Song, J., 103
 Song, S., 79
 Sonis, M., 914, 919, 920
 Sonoiki, D., 1188
 Sonstelie, J., 83, 84
 Sørbye, S.H., 1403, 1413
 Sorger, G., 946, 959, 962
 Soto, A., 762, 780
 Soyster, A.L., 241
 Spaccapietra, S., 712, 713
 Sparrow, F.T., 790, 792, 793, 796
 Speckman, P., 1425
 Spencer, G.M., 1228
 Spiegelhalter, D.J., 1214, 1358, 1407,
 1408, 1448
 Spiekermann, K., 356, 754
 Spielman, S.E., 1321
 Spiro, H.M., 697
 Sprenger, C., 1480, 1482
 Squires, G., 98
 Srholec, M., 384, 385
 Srivastava, R.M., 1469
 Staber, U., 597
 Stampacchia, G., 804
 Stanton, E.A., 1051
 Stark, O., 286
 Steadman, P., 754
 Steel, D., 1167
 Steel, M., 304
 Steenberghen, T., 755
 Stehman, S.V., 1387, 1388
 Stehrer, R., 909
 Stein, A., 1390, 1391, 1393
 Stein, J.L., 20, 23, 910
 Steinke, T., 1376
 Steinmueller, W.E., 380, 385, 386
 Stelder, D., 531
 Stelder, T.M., 894
 Stensgaard, A., 1420
 Stern, H.S., 1408, 1411, 1424, 1576,
 1580–1582, 1584
 Stern, N., 944
 Sternberg, R.J., 323
 Steve, W., 1077
 Stevens, B.H., 888
 Stevens, D., 1386
 Stevens, M.C.G., 1290
 Stevens, P., 342
 Steward, F.R., 1024, 1026
 Stewart, N.F., 770
 Stigler, G.J., 61, 968
 Stiglitz, J.E., 214, 215, 278, 540, 571, 912, 934
 Stimson, R.J., 284
 Stoakes, A., 140, 141
 Stokey, N.L., 962, 963
 Stoll, M., 102
 Stone, L.L., 916
 Stoneman, P., 378
 Storeygard, A., 646
 Storper, M., 444–446, 461, 462, 483, 492,
 494, 496, 595, 600, 601, 605, 650, 653, 654,
 659, 660
 Strange, W.C., 53, 324, 325
 Straszheim, M.R., 695
 Strogatz, S.H., 729, 1267
 Strout, A., 890
 Stubbs, M., 129
 Stumpner, P., 312, 1375
 Sturges, H.A., 1139, 1140

- Sturm, D.M., 584
Stutzer, A., 279
Subramanian, S.V., 1336, 1337, 1341, 1344,
 1345, 1353, 1354, 1356–1358
Südekum, J., 200
Sugihara, K., 1265, 1268
Sui, D., 1120
Sullivan, M., 995
Sun, D., 1425
Sunley, P.J., 448, 459, 466, 468, 470, 593, 594,
 596–598, 604, 611, 616, 617, 626, 668
Suresh, B., 1017
Sutherland, H.,
Svarer, M., 88, 89
Svedin, U., 707, 708, 720
Swait, J., 986, 988
Swan, T.W., 170
Syabri, I., 1153, 1307
Sydow, J., 597, 620, 624
Symanzik, J., 1299, 1300, 1302–1306
Szargut, J., 1024, 1026
- T**
Taaffe, E.J., 726
Tabuchi, T., 543, 550, 552, 554, 567, 582
Tahmisioglu, A.K., 1644
Takahashi, K., 1329
Takayama, T., 789
Takeuchi, H., 478
Takeyama, M., 1224
Talen, E., 1076
Talens Peiró, L., 1022, 1025
Tanaksaranond, G., 1188
Tango, T., 1329
Tanner, M.A., 1575
Tansley, S., 1109
Taplin, J.H.E., 696
Tappeiner, G., 1567
Tate, N.J., 1469
Tatsiramos, K., 120
Tavlas, G.S., 1540
Tawn, J., 1209, 1428
Taylor, C., 51
Taylor, J., 114, 119, 265
Taylor, L.O., 20, 978, 979, 989, 996
Taylor, M., 149
Taylor, P.J., 1160, 1161, 1254
Teece, D.J., 398
Teixeira, A.A.C., 447
Tel Weel, B., 381
Teller, A.H., 1575, 1577, 1579
Teller, E., 1575, 1577, 1579
- Temple, J., 195, 196, 300, 303, 305
ten Raa, T., 890
Tervo, H., 112
Teulings, C.N., 86
Thabane, L., 1431
Thawornkaiwong, S., 1615
Thayer, M.A., 987, 999
Theodoridis, Y., 713
Thirlwall, A.P., 111, 264, 265, 267, 923
Thisse, J.-F., 222, 540, 543, 548–552, 554, 563,
 564, 566, 582, 839, 847
Thom, R., 916
Thomas, A., 1358, 1448
Thomas, B., 1238, 1240
Thomas, D.S., 4–6, 10, 11, 14
Thomas, E.F., 969
Thomas, I., 101, 103
Thomas, J.J., 1189
Thompson, S.K., 1386
Thompson, W.S., 4
Thompson-Fawcett, M., 1078, 1079
Thomson, A., 1279, 1420
Thomson, B., 143
Thorns, D.C., 148, 149
Thrift, N., 274, 497
Thünen, J.H., 745
Tiebout, C.M., 149, 158, 162, 556, 890, 1042
Tiefelsdorf, M., 1441–1444, 1480–1482,
 1514, 1661
Tierney, L., 1577
Tijssen, R.J.W., 493
Timmermans, H.J.P., 718, 720, 722
Timmins, C., 1042
Timonin, V., 1182
Timothy, D., 552
Tingley, M., 1433
Tl, S., 1412
Tobias, J.L., 1661
Tobin, L.T., 246
Tobin, R.T., 238, 244
Tobler, W.R., 737, 1108, 1165, 1188, 1478,
 1481, 1598
Todaro, M.P., 855, 856
Todd, M.J., 244
Tödtling, F., 379, 381–383, 385–387, 459, 461
Tolbert, C.M., 48–51
Tolle, K.M., 1109
Tominski, C., 1152
Torre, A., 498, 499
Torrens, P.M., 1115, 1224, 1228
Train, K., 1629
Trainer, G.A., 888
Trajtenberg, M., 415, 419, 429

- Tranmer, M., 283
 Travers, J., 728
 Travis, J., 1403
 Trefler, D., 909
 Trichtl, G., 51
 Trippl, M., 459, 461, 663
 Trognon, A., 1657
 Troitzsch, K.G., 1238
 Trotter, P., 1164
 Truscott, T., 731
 Tsur, Y., 934, 936, 937, 940, 942, 944
 Tsutakawa, R., 1425
 Tucker, A.W., 764
 Tufte, E.R., 1141
 Tukey, J.W., 1141, 1147, 1296, 1297, 1308
 Tura, T., 386, 460, 463
 Turnbull, B.W., 1320
 Turner, M.A., 645, 646, 677
 Turnovsky, S.J., 969
 Twigg, L., 1238, 1246, 1344, 1345, 1357
- U**
 Uberti, T.E., 660, 661
 Uchida, T., 1444
 Udomsri, R., 749
 Ulam, S., 1227, 1575
 Ullman, E.L., 322
 Ulph, A., 952, 954, 955, 957, 958
 Unruh, G., 466
 Unwin, A., 1307
 Unwin, D.J., 1265
 Usai, S., 660, 662, 663
 Utzinger, J., 1420
 Uyarra, E., 387, 657
- V**
 Vale, M., 380, 386
 Van Alstyne, M., 721
 Van de Weghe, N., 715, 1230
 Van den Berg, G.J., 82
 Van der Linde, A., 1214, 1408
 van der Ploeg, F., 963
 Van der Ryn, S., 1073
 Van der Straeten, K., 282
 van der Voet, E., 1013
 van Dijk, J., 894
 van Dyk, D.A., 1584–1586
 van Ginneken, L.P.P.P., 1243
 van Groenigen, J.W., 1390, 1391, 1394
 van Lierop, W.F.J., 150, 151
- Van Marrewijk, C., 569, 572, 581, 582, 907, 909, 912
 van Mourik, H., 718, 720
 Van Ommeren, J., 82
 van Oor, F., 660–663
 van Oort, F.G., 467, 482, 485, 535, 593, 594, 597
 van Veen, F., 1480, 1482
 Van Vuuren, A., 89
 Van Wey, L.K., 1097–1099
 van Ypersele, T., 89
 Vangenot, C., 713
 Vannahme, M., 750
 Vapnik, V., 1182
 Vardas, G., 946
 Varga, A., 209, 221, 405, 500, 501
 Varin, C., 1630
 Veblen, T., 273
 Vega-Redondo, F., 824
 Venables, A.J., 222, 223, 233, 334, 336, 337, 356, 483, 528, 530, 531, 541, 542, 559, 576, 582, 583, 595, 653, 654, 659, 839, 1367
 Ver Hoef, J.M., 1292
 Verburg, T., 467, 484–486, 535, 593, 594, 597
 Verdoorn, P.J., 265, 268, 271
 Verganti, R., 464
 Vergne, J., 627
 Verhoef, E., 688
 Vernon, R., 320, 397, 404
 Verspagen, B., 357, 381, 384, 385
 Vespiagnani, A., 824, 825, 1271
 Vickerman, R., 69, 356, 726, 738
 Vijverberg, W.P.M., 1556, 1627–1629
 Villalba Méndez, G., 1018, 1019, 1022, 1025, 1026
 Vincent, A., 1323
 Viscusi, K., 1040
 Vitousek, P.M., 1086
 Voas, D., 1240
 Voith, R., 27, 160
 von Ehrlich, M., 357
 von Hippel, E., 402
 von Neumann, J., 1220, 1221, 1227
 von Thünen, J.H., 511, 512, 745
 Voss, H., 1299
 Vounatsou, P., 1292, 1420
- W**
 Wachs, M., 701
 Wackernagel, H., 1462, 1470
 Waddell, P., 749

- Wadsworth, J., 116
Wagener, F., 937
Wagner, G.A., 360, 363, 364, 366
Wagner, P., 753
Wago, H., 1429, 1572
Wakefield, J., 1420
Wakolbinger, T., 799
Walde, J., 1567
Waldman, D.W., 24, 1005
Waldrop, M.M., 1219
Walker, J., 1626
Walker, R., 494
Wall, M.M., 1214
Waller, L.A., 1288, 1328, 1426, 1443, 1444
Wallin, J.F., 1301
Walsh, R.P., 984, 1042, 1043
Wand, M.P., 1408, 1409
Wang, H., 1020
Wang, J., 302, 501, 594, 600, 1177–1180,
 1197, 1279, 1285, 1298
Wang, X., 1301, 1512, 1572, 1627–1629, 1632
Wang, Y.-Q., 286, 968
Ward, M.H., 1449
Wardrop, J., 765
Wardrop, J.G., 790, 792, 795
Warrick, A.W., 1391
Wasserman, S., 727, 728, 734
Wasserman, W., 1443
Wassermann, S., 1263
Waterman, P.D., 1336
Waters, N.M., 727, 735, 737
Waters, W.G., 693
Watts, D.J., 729, 730, 1267, 1272
Waugh, F.W., 994
Weaver, W., 814, 816
Weber, A., 521
Webster, C.J., 1223
Webster, R., 1290, 1392, 1393, 1464,
 1465, 1473
Wegener, M., 356, 743, 744, 749, 750,
 753–755, 838, 849
Wei, Y.H.D., 1625
Weibull, J.W., 744
Weick, K., 469
Weil, D., 321
Weil, D.N., 292
Weil, D.V., 181
Weinberg, B., 100
Weinberg, D., 82
Weinstein, D.E., 548, 554, 576, 583, 584, 909
Weiser, M., 1254
Weiss, S.F., 745
Weissman, J.S., 1314
Weitzman, M.L., 939, 940, 946, 1051
Westlund, H., 535
Wettschereck, D., 1299
Whalley, S., 1168
Wheaton, W.C., 83, 129–131, 552
Wheeler, C.H., 637
Wheeler, D.C., 1426, 1441–1444, 1446–1449
Wheelock, D.C., 360, 363, 364, 366
White, D., 288
White, H.C., 1264
White, M.J., 79
White, N., 1081
White, R., 834, 838, 1223
Whitehead, C.M.E., 138, 139
Whitehead, J., 988
Whitten, S.M., 1033
Whittle, P., 1280, 1281, 1291, 1606
Wie, B.W., 246
Wiener, N., 1220, 1221
Wikle, C.K., 1284, 1292, 1366, 1369
Wildasin, D.E., 969
Wilde, L.L., 64
Wilkie, D., 1164
Wilkinson, L., 1139, 1140, 1145
Williams, H.C.W.L., 778
Williams, K., 1033
Williams, S., 712
Williamson, J., 1381
Williamson, P., 1236, 1240
Willinger, W., 729
Wills, G., 1307
Willumsen, L.G., 761, 769
Wilson, A.G., 744, 751, 762, 776, 778, 815,
 890, 1114, 1500, 1654
Wilson, G., 1394
Wilson, J., 968, 1165
Wilson, M.L., 990, 1133
Wilson, R.J., 1255
Windle, J., 1033
Winer, S.L., 120
Wingo, L., 511
Winkelmann, R., 279
Winstanley, A., 148, 149
Winsten, C.B., 760, 766, 779, 789, 792, 793,
 795, 807
Winter, S., 417, 715
Wirl, F., 939, 945
Withagen, C., 963
Withey, S.B., 282
Witlox, F., 715, 720, 1230, 1254
Witt, U., 625, 627
Wojan, T., 324
Wolff, E.N., 408, 409

- Wolfram, S., 1127, 1222
 Wong, D.W.S., 1158, 1162–1166
 Wong, W., 1575
 Wood, G., 89, 140, 141
 Wood, J., 1187
 Wood, S., 1452, 1453
 Woodward, A.J., 1344
 Woodward, D., 533, 640
 Worboys, M.F., 1108
 Wozniak, A., 12
 Wright, D.J., 1116
 Wright, J., 984, 987
 Wu, B.M., 1178, 1180, 1239, 1251
 Wu, F., 1223, 1224
 Wu, H., 1125
 Wu, J.H., 341, 780, 1044
- X**
 Xepapadeas, A., 940, 942, 946
 Xi, J., 1020
 Xian, G., 1369
 Xiaodong, W., 1021
 Xiaoning, W., 1020
 Xie, J., 1125
 Xu, J., 1033
 Xu, T., 103, 1021, 1023
- Y**
 Yamamoto, K., 233, 234
 Yamarik, S., 361
 Yang, C., 1125
 Yang, H., 1020
 Yang, L., 1369
 Yang, W., 1035, 1038
 Yellen, J., 1255
 Yeung, D., 939
 Yfantis, E.A., 1391, 1393
 Yi, K.M., 913, 918
 Yilmaz, S., 360, 363, 364
 Yin, R., 942
 Yin, Z.-C., 715
 Yinger, J., 97
 Yong, J.S., 694
 Yoo, E., 1321
- Yoon, M., 1526, 1528
 Yorizane, S., 980
 York, J., 1409, 1410, 1413, 1415
 York, R., 1094
 Young, A., 483
 Young, M., 99
 Yu, D., 1222, 1442
 Yu, H., 716, 1020
 Yu, J., 1587, 1588, 1612, 1643, 1644, 1649
 Yu, M., 791, 807
 Yuan, Y., 699
 Yule, G.U., 1159
 Yuzefovich, E., 1611
- Z**
 Zaccour, G., 952, 959, 970
 Zahavi, Y., 746
 Zaitseva, G., 840, 841
 Zax, J., 101
 Zelinsky, W., 1095, 1099
 Zelli, R., 283
 Zellner, A., 999
 Zellner, M., 1228
 Zemel, A., 934, 936, 937, 939, 942, 944
 Zenou, Y., 95, 98, 99, 101, 102
 Zhang, A., 246
 Zhang, D., 246, 790, 791, 808
 Zhang, Q., 646
 Zhang, W.X., 1442, 1444
 Zhang, Y., 1301, 1567
 Zhen, Z., 1020
 Zhou, X., 699
 Zhu, L., 1443, 1444
 Zhu, S., 1490, 1548, 1561
 Zielinski, W., 1420
 Zimbalist, A., 337
 Zimmermann, A.C., 280
 Zipf, G.K., 743, 815, 817
 Zodrow, G.R., 968, 969
 Zografos, G., 755
 Zou, D., 1564
 Zukauskaite, E., 379, 381–383, 385–387
 Zuniga, R., 1058
 Zweig, J.S., 280, 281, 284, 285
 Zylberberg, A., 37

Subject Index

A

ABA. *See* Activity-based analysis (ABA)
Abatement games, 952, 954
ABM. *See* Agent-based modeling (ABM)
Absolute advantages, 905–907
Absolute β -convergence, 299–300, 307, 308
Absorptive capacity, 208, 210, 379, 386, 392, 395, 396, 408, 417, 425, 427, 429, 433, 434, 436, 463–465, 467, 486, 656
Accessibility, 396, 402–405, 743, 745, 748–751, 753, 756, 823–824, 826, 828–829
Accessibility to
employment centers, 80, 81, 90
knowledge, 402–405, 427
product and factor markets, 115
Access to screening, care and treatment, 1321
Accountability, 369
Action, 592–594, 598–605
Activity
implementation, 717
pattern, 707–710, 718, 722
scheduling, 707, 708, 717, 719, 720, 723
space, 708
Activity-based, 763
Activity-based analysis (ABA), 705–723
Activity-based approach, 690, 706–708, 710
Activity-based models, 689, 691
Actors, 490–497, 499–502
Actual spending vs. some proxy, 361
Adaptability, 625, 626
Adaptation, 611, 617–618, 620, 621, 623, 624, 627, 937, 938
Adaptive-cycle model, 598
Adjacency, 1256, 1259, 1268–1270
matrix, 1268–1270
Adjacent possible, 469, 471
Adjoint variables, 249–250

Adjustment
cost, 865, 867, 868
path, 286
Advantages, 476–477, 479, 483, 484
Age, 111, 116–118, 122
Agency, 619
Agent, 1115, 1118
Agent-based modeling (ABM), 502, 717, 719, 720, 723, 750, 752–753, 830, 837, 1115, 1124, 1126, 1128, 1192, 1217–1231
Age structure, 1094
Agglomeration, 76, 85–88, 199–204, 335, 337, 461, 470, 476–477, 479, 483, 484, 541, 542, 545–552, 554–556, 558–560, 563–566, 651, 663–664, 854, 869–871
economies, 36, 53, 56, 148, 266, 383, 403, 409, 421, 422, 432, 476, 477, 485, 508, 509, 513, 520–521, 524–525, 529, 531, 532, 549, 551, 563, 566, 614–616, 632, 634, 636–638, 640, 642–644, 651, 829–830, 835, 1041, 1043, 1049
effects, 632–639, 642, 918
externalities, 335, 477, 485–487
forces, 441
Aggregate behavior, 1219, 1227, 1230
Aggregate place-to-place migration, 6
Aggregation, 1157–1170
Aggregative fallacy, 1345
Aging population, 24
Agriculture, 1050, 1053–1055, 1057–1065
AIC. *See* Akaike information criterion (AIC)
Airline, 1256, 1267, 1270–1271
Akaike information criterion (AIC), 1439–1442
Allocation of land, 511, 525
All-or-nothing assignment, 769
Alternative segments, 781–782

- Alternative specifications, 999
 Ambient awareness, 733
 Ambiguity, 21–22, 26, 32
 Amenities, 19, 21–33, 333–341, 345, 350, 479, 855, 994, 997–1008
 cities, 326
 compensation, 24, 26
 levels, 19, 27, 30, 31
 Amenity-adjusted net wages, 336, 341, 350
 Analytical complexity, 822
 Animation, 1152–1153
 ANNs. *See* Artificial neural networks (ANNs)
 Anthropogenic activities, 930
 Applications of the hedonic method, 148, 158
 Applied science, 1109
 Arc, 1256, 1260
 Area data, 1198, 1279–1281, 1287–1288, 1437
 Areas expected to grow, 27
 Area-to-point Kriging, 1471–1472, 1474
 Articulation of discourses, 464
 Artificial neural networks (ANNs), 1174, 1181–1183, 1190
 Aspatial oligopoly, 239–244
 Aspirations, 279–281, 284
 Assigning traffic, 688, 689
 Associational density, 1548
 Asymmetric, 761, 769, 784
 Attribute theory of demand, 689
 Autarky equilibrium, 905, 907
 Autocatalytic/self-reinforcing mechanisms, 614
 Autocorrelation, 1176–1178, 1180, 1190,
 1191, 1386, 1388, 1389, 1391
 Autocorrelation in health events, 1311–1333
 Automobile, 22–23, 76, 84, 379, 686, 690, 694,
 696, 698, 709, 717, 742, 920, 994,
 1001, 1004, 1065, 1076–1077, 1080
 “Average” location, 21
- B**
 Background variation, 1325
 Backstop technologies, 941
 Backward
 bending supply, 39, 40
 causal impact, 882
 linkage effect, 869
 linkages, 898–899
 Balancing factor, 862
 Barriers to labor mobility, 118–119
 Baseline conditions, 1056, 1058, 1065–1067
 Bases, 379, 380, 385–387
 Basic research, 651, 652, 656, 663
 Basic sector, 265
- Bayesian, 1291, 1292, 1347, 1360, 1447–1449,
 1458, 1556, 1620, 1627, 1628,
 1631–1633
 hierarchical modelling, 1286, 1291–1292,
 1447
 inference, 1196, 1197, 1201, 1209,
 1211, 1213–1215, 1291, 1402,
 1403, 1407–1408, 1428, 1433,
 1573–1578, 1582
 MCMC estimation, 1538, 1545,
 1571–1594, 1598, 1633, 1649, 1660
 methods, 1292, 1402, 1406, 1409
 spatial analysis, 1195–1215
 spatial hierarchical modelling, 1280, 1289
 Behavior, 1236–1238, 1246, 1248, 1250, 1251
 Behavioral
 principles, 792, 808
 rules, 838, 848
 Benchmark specification, 1499, 1573
 Benefit-cost analysis, 974, 975, 987, 989–991
 “Benefit” of living, 22
 Benefits, 12
 Benefit transfer, 975, 987, 989–991
 Benefit transfer method, 975, 989–990
 Best linear unbiased predictor, 1468
 Beta-convergence model, 359, 360
 Betweenness, 1262, 1263
 Between variance, 394
 Beveridge curve, 67–68
 Bid rent, 77, 82–84
 Big data, 721–723
 Binary response, 1621, 1623
 Binomial co-kriging, 1290
 Biodiversity, 976, 1054, 1061, 1075, 1101
 Biological responses, 1055
 Biophysical systems, 1030
 Bioterrorism, 1313, 1327
 Black box, 1109
 Block-permutation matrices, 919
 Bonferroni adjustment, 1490, 1493, 1495
 Boom and bust, 37, 52–56, 140, 153
 Booms and busts/boom and bust, 37, 52–56
 Border effect, 922
 Borders approach, 148, 160–162
 Borrowing of strength, 1425
 Boserup hypothesis, 1090
 Box and whiskers, 1141
 Box plots, 1139–1141
 Braess paradox, 789, 793, 799–801
 Breakpoint, 871
 Breast cancer, 1317, 1321, 1324
 Bridging, 463–464, 469
 Broadband, 335, 342

- Brushing, 1286, 1296, 1297, 1302–1306
Budget, 1386, 1387, 1397
Budget line, 38, 39
Buyer characteristics, 162
Buzz, 379, 483, 595, 596, 654, 659
Buzz-and-pipeline model, 594–596
- C**
California Coastal Commission, 30, 32
Cancer, 1449
Canonical model, 612, 618, 619, 627, 931, 933, 940, 943
Capabilities, 396, 398, 399, 410–411
Capital externalities, 319, 322, 326
Capital income tax, 968, 969
Capitalism, 480
Capitalization
 of environmental quality, 162
 rate, 27, 28, 130, 131
 of taxes, 148, 158, 164
Capital mobility, 852, 868
Carbon lock-in, 466
Car industry, 699
Cartograms, 1287
Cartography, 1138, 1145, 1154
Car use, 693–694, 696, 697
Case attractor, 1330
Case-control studies, 1313, 1329, 1330
Casual ecologic inferences, 1343
Catastrophic events, 930, 937, 942
‘Catching up’, 298, 299
Causal effect, 633
Causal inferences, 1337, 1342, 1361
Causality, 340–342, 478, 480, 486, 487, 669, 672–680
CCF. *See* Cross-correlation function (CCF)
CEAs. *See* Component economic areas (CEAs)
Cellular-automata (CA) models, 285, 752, 830, 834, 836–838, 841, 844, 848, 849, 1112, 1114, 1115, 1124, 1126–1128, 1192, 1217–1231, 1369, 1382
Censored, 1620–1624
Census of population, 1466, 1469, 1471, 1474
Center-periphery model, 869, 870
Central authority, 953, 967, 968
Centrality, 1254, 1255, 1262–1264, 1270
Central limit theorems (CLTs), 1598, 1603, 1614, 1615
 for linear quadratic forms, 1603, 1614
Centrifugal force, 869
Centripetal force, 869–871
- Certainty equivalent, 332, 342, 343, 345, 346, 348, 350
CGE. *See* Computable general equilibrium (CGE)
Change of support, 1470–1472
Changes in stocks, 877
Chaotic behavior, 819–820, 822, 823, 830
Checkerboard, 1387
Chemical industry, 1010, 1013, 1015–1019
Chemicals, 1010–1019, 1026
Chicago, 762, 766, 769, 779–782
Choice experiments, 975, 977, 983, 985–987, 990
Choropleth maps, 1143–1147, 1152, 1298–1299
CIS. *See* Community Innovation Surveys (CIS)
Cities, 475–487, 528–530, 532, 533, 537, 632–647
City structure, 94
Classical location theory, 318, 509, 835
Classification, 1146, 1147
Clean locations, 997, 998, 1006, 1007
Clean places, 1005–1006
Cliff-Ord type models, 1598, 1606, 1611, 1612, 1615
Climate
 amenities, 23
 change, 471, 566, 717, 754, 756, 939–941, 944, 946, 947, 1049–1069, 1101
 and residential mobility, 154
CLTs. *See* Central limit theorems (CLTs)
Cluster-based development, 923
Clustering, 872, 1175, 1176, 1183–1185, 1190
Clustering coefficient, 1258, 1265, 1267
Clustering methods, 713, 1328–1332
Clustering of
 data, 1329, 1348
 entrepreneurs, 207
 firms, 87, 404
 minorities, 98
Cluster-life cycle, 596–599, 605
Clusters, 395, 403–404, 409, 459, 460, 462–465, 469–472, 492, 496, 501, 593–605, 650, 653, 657, 661–664
detection, 1286, 1290
dimension, 594, 604
dynamic, 594–604
emergence, 470
evolution, 594–600, 602–605
of industries, 880
methods, 49–52
CML. *See* Composite marginal likelihood (CML)
Cobb-Douglas, 359, 366

- Cobb-Douglas production function, 172, 179, 180
 Co-created, 468, 470
 Codified, 495, 496
 Codified knowledge, 477, 478, 482, 483, 495–498, 502
 Coefficient of spatial friction, 7
 Co-evolution, 458, 465–471
 Coevolutionary theory, 1327
 Cognitive
 competencies, 477
 dissonance, 462
 proximity, 499, 656, 658, 659, 662, 663
 space, 664
 Collaboration networks, 482
 Collective learning process, 446, 449, 451, 452, 492, 498
 Collinearity, 1442–1446, 1452, 1454–1558
 Color map, 1143, 1145–1147, 1150
 Combinatorics, 1387, 1391, 1393, 1394, 1396
 Combined mode and route choice, 775
 Combined model, 760–763, 778–781, 784
 Commercialization, 205, 206, 208–210
 Commodity flow, 890, 921, 1505, 1654
 Common ecologic effects, 1340–1342, 1347, 1356
 Common factor test, 1526, 1529
 Common property, 1096–1098
 Communication costs, 543, 552–554
 Community detection, 1263, 1264
 Community Innovation Surveys (CIS), 392, 398
 Commuter corridors, 688
 Commuting, 37, 45, 47–51, 54, 55
 behavior, 51, 529, 1228, 1551
 cost, 76–79, 82, 83, 88, 98, 101, 335, 510, 542–551, 553, 555–556, 560, 562, 565, 871
 costs, 76–79, 82–84, 88, 98, 101, 542–551, 553, 555–556, 560, 562, 564, 565
 patterns of women, 102
 Comorbidity, 1316, 1326
 Compact cities, 544, 564–566
 Comparative advantages, 905–907
 Compare the different techniques, 163
 Compensating differentials approach, 154
 Compensation, 22–28, 30, 32
 Compensation shares, 1004
 Competing, 821–822
 causes of death, 1316, 1326
 factor, 854, 856
 Competition, 441, 443–445, 447–449, 451–454
 Competition effect, 869, 870
 Competitive behavior, 788, 1033
 Competitive land and labor markets, 30
 Competitiveness, 617, 621, 622, 624, 627
 Complementarity slackness condition, 765, 771
 Complementary factor, 854, 856, 860
 Complete spatial, 1314, 1323, 1326
 Complex
 adaptive systems, 458, 468, 469
 behavior, 814–815, 820–822, 830
 dynamics, 261
 motions, 819, 822
 system, 814, 821, 823, 825, 829
 Complex co-evolving systems, 467
 Complexity, 459, 460, 465, 467–472, 811–831, 1085, 1086, 1088, 1095, 1099, 1101–1102, 1219–1220, 1222, 1223, 1225–1226, 1228, 1230
 science, 458, 465, 468, 469, 473
 theory, 462–463, 468, 470, 473, 750, 812, 1191, 1219, 1223
 Component economic areas (CEAs), 49
 Composite cost, 775, 777
 Composite marginal likelihood (CML), 1620, 1629–1631
 Compositional, 1342–1345, 1347, 1357, 1358, 1361
 Compositional effects, 283
 Compression, 1110, 1112
 Computable general equilibrium (CGE), 896
 Computational
 purposes, 1390, 1395
 statistics, 1126, 1132
 transportation science, 711
 Computational-process models (CPMs), 717, 719, 723
 Concentrated log-likelihood, 1562–1563
 Conditional autoregressive (CAR) models, 1204, 1213, 1214, 1286, 1359, 1360, 1406, 1409, 1410, 1424, 1426, 1430, 1432
 Conditional β -convergence, 299–303
 Conditional predictive ordinate (CPO), 1407, 1408
 Conditional probability (Monte Carlo simulation), 1236, 1240–1245
 Conditional simulation, 1469, 1470
 Conditional spatial autoregressive (CAR) model, 1281
 Conditional tests, 1523–1524, 1527–1528, 1530
 Conditioned choropleth maps (CCmaps), 1301–1302
 Condition index, 1444–1446, 1454, 1455
 Configurations, 1386–1388, 1390–1393
 Congestion, 335–338, 484, 687–688

- Connections, 727–730, 734, 736–738
Connectivity, 469, 812–814, 818, 820, 823–830
Connectivity degree distribution, 826
Connectivity infrastructure, 829
Conservation, 1030–1033, 1040–1042,
 1044, 1045
Conservation policies, 1031, 1040, 1041
Consistency, 1478, 1495, 1497, 1503
Constant lot size, 29
Constant returns to scale, 293
Constructed regional advantage, 386
Consumer cities, 84, 478
Consumption, 37–40, 478
 capital asset pricing model, 345
 demand coefficient, 892–894
 package coefficients, 892
 possibility frontier, 906–908
Contact generating function, 86
Contagion, 268
Context, 278, 279, 281–286, 592–594, 597–605
Contextual, 1336–1345, 1347–1349, 1351,
 1353–1361
Contextual effects, 283, 1340–1344, 1354,
 1355, 1357, 1361
Contingency, 593, 599, 600, 605, 619, 624
Contingent valuation, 975, 977, 983–988, 1037
Contingent valuation method, 975, 983–985,
 987, 988
Continuous sampling, 1392
Continuous space, 486, 487
Contracted migration, 69, 70
Control group, 676
Convention analysis, 462
Conventions, 459, 460, 462, 464, 465, 470
Convergence, 176–177, 179–181, 184,
 188–190, 278, 281, 354, 357,
 358, 361
 algorithm, 779
 clubs, 307–312
 σ -Convergence, 308–309, 312
Conversion, 618, 619, 621, 622, 625
Converting rents into property values, 27
Convex optimization, 764, 780
Cooperation, 444, 447–449, 452–454
Coordinated market economies, 460
Coordinated multiple views, 1147–1148,
 1150, 1151
Coordination, 261, 262, 271, 274
Core-periphery model, 221, 233
Core-periphery structure, 540, 542, 544, 551
Cornucopian, 1090–1092
Corporate governance, 271
Corporate strategy, 447, 448
Correlation, 1436–1438, 1443, 1444, 1447,
 1448, 1452–1455, 1457
Cost-of-living, 22, 23, 1002
Cost performance functions, 763, 766
Cost-push IO price model, 881, 899
Cost shares, 881
Costs of movement, 20, 22, 32
Cost structure, 877, 884
Counterfactual, 1040, 1041
Covariance, 1389–1391
Covariation, 1312, 1313, 1315–1317, 1320,
 1321, 1325, 1327, 1329, 1393–1395,
 1436, 1437, 1441–1445, 1447–1448,
 1451–1453
Covariogram, 1389–1391
Coverage, 1387, 1390, 1391
Cox process, 1412
CPMs. *See* Computational-process models
 (CPMs)
CPO. *See* Conditional predictive ordinate (CPO)
Creative class, 203, 317–328, 476
Creative destruction, 215, 262, 376, 434, 534,
 616, 620, 627, 632
Creative industries, 321, 323, 326
Creative processes, 323, 325
Creativity, 476–484, 486
Credit market liberalization, 140, 141
Croplands, 1059
Cross-classified structures, 1339
Cross-correlation function (CCF), 1176–1179
Cross-country studies, 292
Cross-elasticity, 693, 697
Cross-sectional studies, 696
Cross-section spatial regression models,
 1511–1532, 1538, 1539
Cross-validation, 1439–1442, 1444
Crowding effect, 221–222, 226–228, 550, 869
Crowding-out, 5, 12–14, 358
Cumulative causation, 268–269, 271, 869
Cumulative indirect effects, 1540, 1543, 1544
Curve-fitting, 1436
Customer networks, 493
Customer–seller interactions, 839
Cuzick and Edwards test, 1329, 1331
CyberGIS, 1117–1118
Cyberinfrastructure, 1117
Cycle, 760, 763, 764, 773–775
Cyclical economic activity, 11

D

- Damages that are *perceived*, 1006
Damages that are *unperceived*, 1006

- Dams, 342
 Data, 477, 484–485
 generating process, 1554, 1555, 1559
 mining, 716, 719, 720, 1153–1154,
 1173–1174, 1190–1192, 1233,
 1277, 1280, 1288–1290, 1299
 models, 1312, 1325, 1328–1332
 quality, 1292
 Deadweight loss, 46
 The Death of distance, 726
 Decentralization of jobs, 551–554
 Decision-making, 788, 790–801, 808
 Decision-support tool, 369
 Decision tree, 343, 348
 Definition, 1255
 Deforestation, 1030–1031, 1040–1041, 1062
 Degree, 1255, 1257–1259, 1268–1270
 distribution, 825, 1257, 1268, 1270
 number, 729
 Delaunay triangulation, 1259, 1265, 1268
 ‘De-locked’, 610, 615, 624, 626
 Demand
 for children, 1096, 1097
 elasticity, 693, 694, 696, 697
 for housing, 1001
 for labor, 21, 22, 26
 for land, 21
 Demand-driven IO quantity model, 876–881
 Demand-driven theories, 259–275
 Demand-led growth models, 261
 Demand-pull IO price model, 900
 Demands for, 1001, 1003–1005
 desirable locations, 23
 Demand-side influences, 23
 Demo-economic model LINE, 896
 Demographic factors, 116–118
 Demographic transition, 1091–1092, 1099
 Dendrogram, 50
 Density, 476, 484–485, 1257–1259, 1268, 1269
 Density function, 309, 310, 312
 Depletion, 932, 934
 Depreciation, 852, 867–868
 Depreciation rate, 131, 137
 Derivative agglomeration benefits, 632,
 635–638, 643, 644
 Derived demand, 43, 114
 Descriptive statistics, 669, 671, 672
 Desegregation strategy, 103
 Design, 461, 463–467, 471, 473
 Design effect, 1338
 Desirable
 areas, 24, 29
 locations, 1001–1003, 1005, 1006
 Destination effects, 1661, 1664–1672
 Determinant of migration, 4, 5, 10, 14
 Determinants of regional disparities,
 111–112, 122
 Deterministic, 819, 1236, 1240–1241,
 1243–1247, 1250
 Deterministic route choice, 765–769, 772, 775
 Deterministic user equilibrium (DUE), 768,
 770, 775, 781
 Developmental-evolutionary model, 620–627
 ‘Developmental-evolutionary’ model of path
 dependence, 620–627
 ‘Developmental-evolutionary’ view of path
 dependence, 612
 Development path, 342, 395, 428, 485, 593,
 616, 622, 626, 848, 1280
 Deviance, 1556, 1557, 1561
 Deviance information criterion (DIC), 1214,
 1408, 1415
 Diagnostic MCMC, 1582, 1583
 Diagnostic tools, 1436–1437, 1443–1446,
 1456–1458
 Diagonalization algorithm, 243–244, 257
 Diameter, 1260
 DIC. *See* Deviance information criterion (DIC)
 Dichotomous choice, 984, 986, 988
 Difference, 467, 468, 470
 “Differential games,” 959
 Differential variational inequality (DVI), 238,
 246, 249–251, 257
 Differentiated and tradable good, 544, 547
 Diffuse test, 1524
 Diffusion, 391–411, 477–483, 486, 650,
 652–654, 657, 659, 661, 662, 664
 Diffusion modeling, 1283
 Digital data, 711
 Digraph, 1256, 1260
 Diminishing returns, 334, 336, 339
 Di Pasquale and Wheaton model, 129, 130,
 132–134
 Direct, 1561
 effects, 364–365, 1538, 1540, 1541,
 1543–1546, 1550
 representation, 1519–1520
 Directed, 1254–1256, 1258–1260, 1262, 1269
 Discounting, 931–933, 936, 937, 939–940,
 942, 944
 Discount rate, 345
 Discouraged workers, 42–43, 45
 Discrete choice models, 689, 815
 Discrete responses, 1620–1623, 1625,
 1628, 1632
 Discretization, 1387, 1392, 1395

- Disease clusters, 1279
Disease mapping, 1403, 1407, 1409–1411
Disease vectors, 1067
Disequilibrium explanation, 111
Disinvestment, 852
Disorganized complexity, 814, 819, 828
Dispersion, 308, 309, 541, 542, 547, 549–552, 554, 556, 563–566, 770–771, 775–780
Disseminating information on jobs, 96, 103, 105
Distance, 476, 479, 481, 486, 1256, 1257, 1259–1262, 1265–1268
friction, 812, 817
transforms, 1133
Distribution
approach, 292, 308–313
dynamics, 292, 309, 312–313
Diverse knowledge bases, 463
Diversified labor market, 86
Divide-and-conquer, 1112
Division of labor, 483
Dominant design, 466
(Doubly-constrained) SIM, 816
Downscaling, 1114
Dramaturgy, 464
Drinking water, 1066
Droughts, 1053, 1055, 1058, 1060, 1064
Dual graphs, 1259, 1268–1270
Dual variable, 249
DUE. *See* Deterministic user equilibrium (DUE)
DVI. *See* Differential variational inequality (DVI)
Dynamics, 812–815, 818–824, 826, 828–831
behavior, 813, 814, 820, 824, 825, 828–829
brushing, 1286–1287
complexity, 812–816, 818–823, 827, 830
entropy, 820, 822
games, 946, 952, 959–963, 969–970
interactivity, 1286–1287
of location, 834, 836
microsimulation, 1237–1240
models, 742, 751
network oligopoly, 238, 246–257
systems, 815, 819–820
- E**
Earnings
disparity, 36
functions, 26
Easterlin paradox, 281, 285
Eco-innovation, 466, 471
- Ecological
analysis, 1292
effects, 1340–1342, 1344–1345, 1347, 1352, 1354, 1356–1358
fallacy, 1110, 1360, 1361
Ecologic inference, 1335–1361
Econometric models, 717–718, 723
Economic governance, 273–274
Economics
base multiplier, 265
efficiency, 261
geography, 649–664
growth, 292
of happiness, 279, 280, 287
opportunity, 4–7, 12, 14, 15
Economies, 476–480, 482–487
Economies of scale, 149–150, 869
Ecosystem, 1049, 1050, 1052–1055, 1057, 1060, 1061, 1064
Ecosystem resilience, 1055
Ecotones, 1206
EDA. *See* Exploratory data analysis (EDA)
Edge, 1255–1271
The Edge of chaos, 469, 471, 473
EEG. *See* Evolutionary economic geography (EEG)
Effectiveness, 356
Effect on migration, 120
Effects of UI, 119, 120
Efficiency, 857, 860, 871, 1494, 1495, 1497
Efficiency wage, 855
Efficient, 354, 356–358, 369
Eigenfunctions, 1478, 1481, 1482, 1486–1488, 1496
Eigenvalues, 311, 1410, 1421, 1422, 1424, 1480–1482, 1485, 1487, 1488, 1496, 1515, 1538, 1539, 1559, 1564–1567, 1574, 1660
Eigenvectors, 1481–1500, 1502
Eigenvector spatial filtering, 1481–1491, 1494–1500, 1502, 1503, 1505
Einstein's law, 818
EKC. *See* Environmental Kuznets Curve (EKC)
Elastic demand, 790, 801, 802, 805–808
Elasticity, 43, 44, 46, 47
of public capital, 358, 359
of substitution, 858, 859
Elasticity parameter, 817, 827
Electronic health records, 1313, 1314
Embeddedness, 273, 496, 594, 658, 662
Embodied, 407–409
Emergence, 458, 459, 461, 465, 466, 469–472, 812, 824, 825, 827–830

- Emergence theory, 470
 Emergent properties, 1114
 Emission
 coefficients, 880
 games, 952–957, 959
 Employment, 4, 5, 7, 9–14, 333, 334, 343
 growth, 5, 6
 multipliers, 880
 rates, 12, 55, 113–117, 121
 Endogeneity, 302–303, 305, 307, 486, 633, 634, 646
 Endogenous
 amenities, 1003, 1005
 disamenities, 24
 economic growth theory, 477
 growth, 170, 180, 333, 339, 350, 414, 426–431, 853
 growth theory, 170, 193–210, 213–235, 355, 366, 414, 478, 853
 innovation, 465
 location, 569–587
 Energy scarcity, 756
 Energy use coefficients, 880
 Enterprise zones, 96, 104, 121, 676, 678, 679
 Entrepreneurial regional innovation systems (ERIS), 458–459
 Entrepreneurs, 478, 480, 483–484
 Entrepreneurship, 194, 195, 205–210, 632, 639–642
 Entrepreneurship theory, 206
 Entropy, 467–468, 816, 817, 827
 function, 816
 maximization, 817, 818, 820
 Environmental
 agreements, 952
 economics, 952, 960, 969–970, 989, 1030, 1031, 1033, 1039, 1044, 1073
 health, 1039
 management, 945, 1290
 performance, 1053
 quality, 157, 162, 996–1000, 1004, 1005, 1007, 1008
 valuation, 975, 991, 996–997, 1004, 1010
 Environmental Kuznets Curve (EKC), 1091, 1092
 “Epidemic” model, 821
 Epidemiological transition, 1089
 Epistemic communities, 654
 Equilateral triangular, 1387, 1391, 1393
 Equilibrium
 conditions, 789, 790, 792, 793, 795–797, 801, 803, 805–807
 models, 750–752
 path dependence, 618
 state, 816, 818, 829
 theory, 279, 281
 unemployment, 111, 112, 114, 119
 view, 23, 27
 Equitable distribution, 261
 Ergodic distribution, 310–311, 313
 Erie Canal, 341, 342
 ERIS. *See* Entrepreneurial regional innovation systems (ERIS)
 Erratic behavior, 814
 ESDA. *See* Exploratory spatial data analysis (ESDA)
 ESTDA. *See* Exploratory space-time data analysis (ESTDA)
 Estimation, 1236–1239, 1241, 1243–1251, 1638–1651
 Estimator, 1387, 1389–1394
 Europe, 650, 661, 663
 Europe 2020, 471
 Evaluate policy alternatives, 159
 Evapotranspiration, 1053
 Evolution, 528, 533–536, 813–815, 818, 822, 824, 829
 Evolutionary, 490, 494, 498, 501–502
 Evolutionary algorithms, 1126
 Evolutionary complexity, 813
 Evolutionary economic geography (EEG), 458, 466–469, 485, 487, 498, 540, 591–605, 610–612, 627, 656
 Excitable, 731
 Exergy, 1010, 1024, 1026
 Existence value, 976, 985, 987, 989
 Exogenous
 lagged variables, 1517, 1520
 uncertainty, 933–938, 945
 Expansion method, 1289
 Experiences, 478, 483
 Experimental
 regionalism, 461, 463
 variogram, 1463–1464, 1466
 Exploitation sub-system, 458
 Exploration sub-system, 458
 Exploratory data analysis (EDA), 1286, 1290, 1296–1297, 1303, 1307, 1308
 Exploratory space-time data analysis (ESTDA), 312, 313
 Exploratory spatial data analysis (ESDA), 292, 307, 312, 313, 1150, 1286–1287, 1290, 1295–1308, 1367
 Explosive growth, 338, 339
 Exponential, 1388–1390, 1395
 Exponential model, 1464, 1465

- Export and import volumes, 1066
Export base theory, 556, 558, 561
Exposure routes, 1316
Extensive income growth, 894
Extensive margins, 922
External economies of scale, 440–443
External effects, 860
Externalities, 483–485, 913, 922, 1087, 1090, 1092, 1096, 1101
Extrapolations, 687, 688, 700
- F**
- Face, 1259, 1260, 1268
FACE experiments, 1053–1054
Face-to-face (FTF) interactions, 400–403, 405, 490, 491, 496, 498, 499
Factor content of trade, 909
Factor mobility, 170, 529, 570, 670, 851–872, 908
Factor movements, 530, 564, 871
Factor price equalization (FPE), 909, 910
Family, 6, 14
FDI. *See* Foreign direct investment (FDI)
FEAs. *See* Functional economic areas (FEAs)
Feasible generalized spatial two-stage least squares (FGS2SLS) estimator, 1611, 1613, 1615
Feasible policy, 931, 936, 944
Feasible set, 798, 803
Feedback
 effects, 1537, 1539, 1541, 1542, 1545, 1546, 1549–1551
 loops, 914, 918–920, 1096, 1098
 Nash equilibrium, 962, 965
 strategies, 960, 964–966
FGS2SLS estimator. *See* Feasible generalized spatial two-stage least squares (FGS2SLS) estimator
Financial
 capital, 852–853, 867
 networks, 788, 789
Finite-dimensional mathematical program, 251
Firm, 638–642
 amenities, 1004, 1008
 performance, 392, 393, 395–400, 407
 renewal, 393–394, 396–398
 routines, 392, 393, 397, 398
 survival, 393
First demographic transition, 1091
First Law of Geography, 737
First round, 879, 888
Fisheries, 1059
Fisher scoring, 1558
Fisher's score function, 1557
Fisheye lenses, 1149
Fitness landscape, 468
Fixed demands, 801–806
Fixed-effects regression model, 1349
Flat, featureless plain, 1001
Flexibility of wages, 111
Flexible specialization, 318, 320, 443, 444
Floods, 1058, 1064
Flow data, 1278, 1279
Focus+context, 1149
Focused test, 1288
Focussed clustering, 1431–1432
Food security, 1055, 1067
Foreign direct investment (FDI), 852, 853
Formal sector, 855–857
FORTRAN, 1487, 1488, 1490
Forward causal impact, 882
Forward linkage effect, 869
Forward linkages, 899
Four-stage model, 688, 689
Fourth Paradigm, 1109
FPE. *See* Factor price equalization (FPE)
Fragmentation of production, 913, 923
Frames, 472
Framework, 888–890
Framing, 459, 463, 465, 472
Free entry and exit of firms, 912
Free location choices, 97
Free trade agreement, 920
Free trade equilibrium, 906, 908, 912
Frequentist's, 1282, 1291
Friends' friends' friends, 731
Fringe benefits, 26, 27
FTF interactions. *See* Face-to-face (FTF) interactions
Fuel price elasticities, 693
Full hedonic equilibrium, 25
Full linearization, 780
Full-mobility equilibrium, 1003
Full value of the environment, 1004
Functional
 form, 996, 998, 999
 specialization, 632, 643
 urban region, 395, 409, 410
Functional economic areas (FEAs), 52, 670
- G**
- Gabriel graph, 1265, 1266
GAM. *See* Generalized additive model (GAM)
Game of Life, 1114

- Gamma distribution, 1199, 1201, 1205
 Gaussian Markov random fields (GMRFs),
 1403–1406, 1409, 1410,
 1412, 1415
 Gaussian model, 1464, 1465
 Gautreaux Program, 104
 Geary's c-test, 1281, 1282
 Gender differences, 117–118
 General equilibrium, 260, 269
 Generalized additive model (GAM), 1452, 1453
 Generalized method of moments (GMM),
 1598–1616, 1628
 General purpose technologies (GPTs), 480
 General-to-specific, 1529
 Genetic algorithms (GAs), 1130, 1131
 Genetics, 1312, 1315–1318, 1327
 Genetic variance, 1317
 Geocomputation, 1108–1110, 1112–1115,
 1118, 1119, 1121, 1123–1135,
 1190, 1222
 Geodesic, 1260
 Geographical
 attractors, 1322
 brushing, 1303
 economics, 570, 585, 667–681
 frictions, 922
 proximity, 467, 471, 490–492,
 498–500, 502
 Geographical information systems (GIS), 1124
 Geographically weighted regression (GWR),
 162, 307, 1134, 1289, 1307,
 1394, 1435–1458, 1523, 1619–1620,
 1624–1625, 1633
 Geographic information, 1110–1112, 1116,
 1117, 1119, 1120
 Geographic information science (GIScience),
 1108–1116, 1118–1121
 Geographic information systems (GIS), 711,
 1286, 1290, 1306–1308, 1360
 Geographic location, 319, 709, 711, 714, 913,
 1116, 1301, 1324
 Geography of innovation, 214, 375–388, 1572
 Geography's quantitative revolution,
 1280–1282, 1284
 Geometric, 1265, 1266, 1268, 1270
 Geometric coverage, 1387, 1390
 Georeferenced data, 1313, 1314
 Geosimulation, 1126–1128
 Geospatial analysis, 1123–1135
 Geostatistical, 1290, 1291
 Geostatistical data, 1297, 1307
 Geostatistical models, 1406, 1411–1412,
 1461–1474
 Geostatistics, 1278, 1280, 1281, 1289–1291,
 1388–1393, 1403, 1406, 1411–1415,
 1420, 1429, 1462, 1463, 1470,
 1473–1474
 Geovisualization, 1137–1154
 German theorists, 726
 Getis and Ord's G_i , 1176
 Getis' G_i , 1626
 Ghettos, 94
 GHG. *See* Greenhouse gases (GHG)
 Ghosh-inverse, 898, 899
 Gibbs sampling, 1575, 1576, 1578–1580,
 1583–1586
 GIS. *See* Geographic information systems (GIS)
 Glaciers, 1058, 1061
 Global
 diffusion, 1539, 1541
 focused and local tests, 1320, 1330
 geography of innovation, 375–388
 and local spillovers, 1536, 1537, 1542,
 1546, 1551
 pipelines, 595, 654, 658–659
 spatial spillovers, 364
 Globalization, 325
 Global pipeline, 496, 497, 595, 654, 659
 GMM. *See* Generalized method of moments
 (GMM)
 GMM estimation, 1598–1616
 GMRFs. *See* Gaussian markov random fields
 (GMRFs)
 Gothenburg model of the Lisbon Strategy, 471
 Governance, 458–461
 GPTs. *See* General purpose technologies (GPTs)
 GPUs. *See* Graphical processing units (GPUs)
 Gradient of subjective wellbeing, 284
 Gradients, 1557, 1558, 1560
 Grand Challenges, 471
 Graph, 1255–1270
 rewiring, 1267
 structure, 1255–1258, 1264, 1268, 1269
 theory, 727, 824, 1253–1257
 Graphical processing units (GPUs), 1118
 Gravity
 equation, 200
 models, 4, 7, 8, 151, 152, 501,
 815, 817, 862–863, 890,
 1654, 1656, 1659,
 1662–1671
 principle, 816
 Greenhouse gases (GHG), 564–566
 Griliches, Z., 994
 Groupwise heteroskedasticity, 1521
 Growth econometrics, 292, 302, 312

- Growth theory, 355–356, 361, 366, 369
Grubel-Lloyd index, 910, 916–918
GWR. *See* Geographically weighted regression (GWR)
- H**
- Halton, 1629
Hamiltonian, 932
Happiness, 278–280, 283–288
Happiness clusters, 283
Harmonic function, 1202
Harrison White, 728
Harris-Todaro hypothesis, 855
Hausman test, 367, 1525, 1541, 1547
Hazard functions, 67
Hazardous waste dumps, 31
Hazard rate, 934–936
HDR. *See* Highest density region (HDR)
Health, 1236, 1237, 1239, 1240, 1245, 1246, 1248, 1250, 1251
Health effects, 30–31, 990, 1006, 1040, 1322
Health event clusters, 1311–1333
Healthy worker, 1316, 1322
Heat island effects, 1052–1053
Heat waves, 1053, 1055, 1058, 1060
Heavy tail, 729
Heckscher-Ohlin model, 907–910
Hedonic, 993–1008
 price indexes, 148, 158, 159
 property method, 975, 977–979, 987
 valuation, 1035–1037
Herd immunity, 1323
Hessian, 1557, 1558, 1560, 1561
Heterogeneity, 392–396, 398, 399, 410, 861, 1030–1034, 1037, 1039, 1041–1044, 1046, 1279, 1280, 1284, 1286–1289
Heterogeneous Poisson model, 1320, 1329
Heterogeneous products, 910, 911
Heteroskedastic, 1279
Heteroskedasticity, 1176, 1352, 1530, 1573–1574, 1585, 1587–1589, 1608–1609, 1612
Heteroskedasticity and autocorrelation-consistent (HAC), 1522
Heteroskedastic spatial regression, 1574
Heuristic techniques, 1397
Hexagonal, 1387
Hexagonal lattice, 1204
Hicksian demand, 40, 83
Hidden innovation, 376
Hierarchical models, 1292
Hierarchical statistical modeling, 1280, 1292
Higher order, 888
Highest density region (HDR), 1201
Highly leveraged homeowners, 89
High-order processes, 1521
High-rent locations, 23, 29
Hindcasting, 1115
Histogram, 1139–1141, 1146, 1149–1151
HME. *See* Home market effect (HME)
Hollowing out, 880, 913, 914, 916, 923
Home market effect (HME), 221–222, 227–228, 232, 570, 573–578, 580, 582, 583, 585
Homogeneity, 21
Homogeneous household/firm, 1001
Homophily, 730
Horizontal dimension, 595
Horizontal intra-industry trade, 916
Horizontal product differentiation, 916
Household-formation events, 149
House price changes, 77, 136
House price hedonics, 148, 155–163
House prices, 1209–1212
House prices and planning, 138
Housing, 333–338, 340–341, 350, 540, 542, 544, 546, 566–567
 affordability, 143
 bust, 163
 demand, 21, 135–138, 140
 density, 25
 discrimination, 97, 101–103
 dissatisfaction, 151
 market prices, 118
 policy, 137, 139
 prices, 148, 151–153, 155–163
 submarkets, 138
 subsidies, 136, 138–139
 supply, 135–138, 337
Housing market
 boom and bust, 140
 and global economic crisis, 142–143
 and macroeconomy, 139–143
Housing supply and development control, 138
Hubs, 825–829
Human, 317–328
 capital, 64, 70–72, 110–111, 114–116, 118, 122, 317–328, 857, 865
 capital externalities, 637
 consequences, 110
Human capital externalities, 319, 322, 326, 637
Human health, 1030, 1031, 1055, 1077, 1088, 1321

- Humanitarian supply chains, 791
 Hydroelectrical, 1059, 1062, 1063, 1065
 Hyperprior, 1204
- I**
- ICAR model. *See* Intrinsic conditional autoregressive (ICAR) model
 “Iceberg” form, 223
 ICTs. *See* Information and communication technologies (ICTs)
 Identical preferences, 24
 Identical production functions, 24
 Identification, 1038–1041
 issue, 160
 strategies, 674–676, 678
 Identifying the demand for a hedonic attribute, 157
 IEA. *See* International environmental agreement (IEA)
 IEA Game, 954, 957–959
 Ignorance uncertainty, 933–936
 Immunity, 1316, 1321–1323
 Impact analysis, 880
 Imperfect
 competition, 910
 substitutability, 1089–1090
 Implicit house price hedonic characteristics, 156–157
 Implicit price, 156–159
 Inada conditions, 171, 186
 In-and out-migration, 6–8, 10, 12
 Incentive policies, 1030
 Incidence matrix, 1269
 Incident cases, 1312, 1320, 1326, 1328, 1329
 Income effect, 40, 189, 283, 558,
 887–888, 1092
 Income elasticity, 83, 137, 260, 264, 268, 275,
 450, 693, 696–697
 Income elasticity of housing demand, 137
 Income multiplier, 880, 893
 Income spillover, 888
 Increasing, 476–484, 486
 returns, 325, 334–335, 337, 476–477, 483,
 484, 540, 541, 544, 547
 returns to scale, 86, 210, 269, 335, 337, 483,
 529, 570, 571, 573, 579, 582, 869,
 871, 910, 1091
 Incremental jobs, 13
 β Index, 1257
 γ Index, 1259
 Indifference curve, 37, 39–41, 905, 908
 Indirect, 1561
 Indirect effects, 365, 1538, 1540, 1541,
 1543–1546, 1550
 Induced feedback effects, 891
 Induced spillover effects, 891
 Industrial clusters, 440–441, 447–455, 479
 Industrial districts, 318, 439–445, 449–455,
 483, 491–494, 536, 594, 603, 616,
 617, 655, 839
 Industrial location, 320, 508, 523, 609, 611,
 614–615, 624, 791, 834
 Industrial location patterns, 610, 611, 615–617,
 624
 Industrial metabolism, 1012
 Industrial mix, 114
 Industry-life cycle, 383, 594, 598
 Inequality, 36, 55, 56
 Inequities in health, 1341
 Infant industry argument, 912
 Infection transmission, 1317, 1318, 1322,
 1323, 1327
 Inferential statistics, 1320, 1328
 Inflationary dynamics, 270
 Influential observations, 1443, 1444
 Informal sector, 856
 Information, 200, 202, 203, 467–469, 471, 472,
 477, 480
 demands, 350
 hurdles, 105
 matrix, 1557–1561, 1563
 Information and communication technologies (ICTs), 476, 495–498, 707, 716
 codified knowledge, 495, 496
 Infrastructure, 331–351, 476, 482, 632,
 645–647
 Infrastructure networks, 1050
 Initial equilibrium, 1001
 INLA. *See* Integrated Nested Laplace Approximation (INLA)
 Inner-cities, 94–98, 103, 105, 543
 Inner-city development strategy, 103
 Innovation ideas, 406, 409–410
 Innovations, 194–200, 203, 205–209, 319,
 323–328, 335, 415, 417–419, 422,
 424–431, 434, 475–487, 533–534,
 593–595, 598, 604, 632, 638–639,
 641–643, 646–647, 650–664
 capabilities, 396, 399
 milieu, 392–393, 396, 399, 402–405, 408–411
 paradox, 465
 platforms, 386
 policies, 481, 482, 652, 654, 657,
 663–664
 and structural change, 616

- Innovative
 capability, 376, 384, 386, 388, 446
 milieu, 439–455, 491, 492, 496
 performance, 650, 651, 655, 657, 659,
 661–663
- Input coefficients, 884–886, 888
- Input-output (IO) linkages, 356
- Institutional approach, 128, 194, 195,
 208–210, 535
- Institutional proximity, 498, 499, 501, 658,
 659, 662, 663
- Institutional regional innovation system
 (IRIS), 458
- Institutional-relational places, 651, 656–657,
 659, 660
- Institutions, 36–37, 48, 54, 194–196, 204,
 208–210, 527–537
- Instrument, 633–636, 640
- Instrumental variable estimator, 1597–1616
- Instrumental variables, 341, 674–675
- Integrated
 approach, 651, 659
 frameworks, 658–663
 IO-spatial econometric approach, 367
 model, 941, 943, 944
- Integrated Nested Laplace Approximation
 (INLA), 1403–1409, 1411–1415
- Intensification, 1090, 1091, 1095, 1097, 1098
- Intensity, 11
- Intensive income growth, 894
- Intensive margins, 922
- Interactions, 490–491, 493–499, 501, 502
- Interaction term, 29
- Intercity trade, 548, 551, 565
- Interconnectivity, 1050, 1053, 1065–1066
- Interdependencies, 285
- Interindustry trade, 910, 916
- Intermediaries, 458, 459, 461, 464, 469, 470
- Internal freshwater resources, 1060, 1063
- Internal migration, 4, 6, 10, 12–14
- International environmental agreement (IEA),
 952, 954, 957–959
- International real estate markets, 134
- International trade, 234, 540–542, 569–587,
 670, 791, 904, 905, 909, 914,
 920–923, 1283, 1655
- Internet, 339, 1254, 1271
- Internet retail purchases, 896
- Interorganizational network, 420, 502
- Interpolation, 1278, 1281, 1436, 1443,
 1455, 1457
- Interpolation autocorrelation, 1315, 1321
- Interpretation of path dependence, 612, 614
- Interregional
 externalities, 360
 feedbacks effects, 880–881,
 886, 887
 flows, 70
 FPE, 910
 income multiplier, 885
 IO quantity model, 884
 Leontief-inverse, 885, 886
 migration, 150, 923
 migration of production factors, 910
 redistribution, 891, 894
 spillover effects, 886–888
 spillovers, 887
 transportation infrastructures, 356
- Interregional input-output (IRIO) model, 366,
 875–900
- Interregional input-output table (IRIOT),
 882–890, 897
- 9-Intersection, 1108, 1109
- Interspecies competition, 821
- Interstate flows, 919, 922
- Intertemporal substitution, 187, 189
- Intraclass correlation, 1347, 1351, 1353
- Intra-industry trade, 910, 914, 916, 918,
 922, 923
- Intra-regional
 effects, 1664–1672
 induced effects, 891
 multipliers, 880, 886, 887
- Intra-urban moves, 150
- Intrinsic conditional autoregressive (ICAR)
 model, 1204, 1286, 1292
- Inverse distance weighting, 1464
- Investment, 332–346, 348–351, 852–853, 857,
 867–869
- IPAT, 1087
- IPCC. *See* The Intergovernmental Panel on
 Climate Change (IPCC)
- IRIO model. *See* Interregional input-output
 (IRIO) model
- IRIOT. *See* Interregional input-output table
 (IRIOT)
- IRIS. *See* Institutional regional innovation
 system (IRIS)
- Irreversibility, 614, 616, 931, 945
- Isosurface, 1175, 1188, 1191
- Isotropic, 1391, 1393
- J**
- “Jacobian spillovers,” 655, 661
- Jacobs externalities, 442

- Jobs, 5, 7, 9, 11–15, 631–647
 accessibility, 81–83, 102
 creation, 359, 632, 644–645
 dynamics, 632
 offer, 61–64, 68, 69
 opportunities, 5, 69, 76, 94, 96, 98, 99, 101, 326
 search, 37, 42–43, 59–72, 79, 85–86, 98–99, 102, 119–120
 costs, 99
 efficiency, 98
 effort, 99
 models, 60–65, 69
- Jobs follow people or people follow jobs, 148, 154
- Join count test, 1281
- Joint product, 686
- Joint test, 1526, 1527, 1530
- K**
- Kaldor-Dixon-Thirlwall model, 265–269
- Kernel bandwidth, 1438–1442, 1445
- Kernel function, 1437–1442, 1445, 1447, 1448
- Kernel ridge regression, 1183
- Keynesian economic management, 269
- Keynesian multiplier, 261, 264, 265
- Keynesian theory, 261, 263, 270
- Keynesian theory of macroeconomic equilibrium, 263
- Knowledge, 490–502, 852–853, 860
 accession, 393, 400, 401, 403, 408
 creation, 196–210
 delivery, 716, 721, 722
 diffusion, 391–411
 economy, 491, 494–497, 502
 exchange, 386, 399–401, 403, 411, 448, 452, 490, 492, 494, 534, 655
 externalities, 356, 366, 382, 413–436, 495, 502
 filter, 380–381
 flows, 401, 413–436, 653–661, 663, 664
 intensity, 392, 393, 395, 396, 398–400, 402–403, 410
 interaction, 396, 400, 401, 403, 405
 production, 320, 325
 sources, 379, 383, 393, 394, 398–401, 404–405, 410–411, 421–422, 655
 spillovers, 215, 218, 221–223, 226–228, 230–235, 396, 399–400, 403, 405, 415, 416, 418–436, 463, 467, 477, 481, 482, 484–486, 495, 496, 500–501
 workers, 320, 476, 477, 481, 486
- Knowledge-based society, 481
- Knowledge-oriented policies (KOP), 387
- Knowledge production function (KPF), 417, 418, 426, 427, 429, 433, 477, 478, 481, 486, 500
- KOP. *See* Knowledge-oriented policies (KOP)
- KPF. *See* Knowledge production function (KPF)
- Kriging, 1208, 1209, 1211–1214, 1463, 1464, 1468–1471, 1473
- ordinary, 1208
- variance, 1390–1396, 1468, 1469
- Kuhn-Tucker conditions, 792, 793, 798–799
- L**
- Labor, 542, 544, 546–550, 555–557, 559–561
 demand, 43–45, 111, 112, 114–116, 119, 122
 mobility, 478, 482–484, 852, 854, 855, 865
 productivity growth, 894
 unions, 36–37
- Labor-augmenting technology, 177–179, 182
- Labor-demand, 43–45, 110–116, 119, 122, 155, 518, 853, 855, 861–863, 1005, 1099
 factors, 111, 114
 shock, 155
- Labor force
 participants, 25
 participation, 36, 41, 53–55, 111, 115, 161
- Labor income coefficient, 892
- Labor market areas (LMAs), 36, 41–43, 47–56
- Labor markets, 17–33
 arbitrage model, 22–23
 areas, 36, 41–43, 47–55, 86
 definition, 47–52
 disequilibrium approach, 23
 equilibrium, 17–33, 35, 45–47, 518
 frictions, 87
 model, 35–56, 155
 rigidities, 36, 43
- ‘Labor-market tightness’, 66, 67
- Labor monopsony, 46
- Labor supply, 36–43, 45, 46, 48, 53–54, 110–111, 113–114, 116–118, 122
 constraints, 115–119
 factors, 111
 and labor demand shocks, 155
 shock, 155
- Labor unions, 36–37
- Ladder-of-life, 288
- Lagrangian multipliers, 39, 796, 799, 1469, 1524–1528, 1530
- Lags, 1388, 1389, 1391

- Land, 542, 544–548, 550, 553, 556, 562, 566–567
Land conservation policies, 1031
Land owner, 855
Land-use, 741–756
 change, 742, 752, 1031, 1041, 1044, 1129, 1218, 1219, 1224, 1228, 1283, 1381
 planning, 743, 1128
 transport feedback cycle, 743, 744, 746
 transport interaction, 741–756
Laplace approximation methods, 1401–1415
Laplace's equation, 1202
Lasso, 1446, 1447
LATs. *See* Location-aware technologies (LATs)
Lattice, 1267, 1270
Lattice data, 1297, 1298, 1305
Law of large numbers, 1598, 1615
Layering process, 618, 619, 621, 622, 625
Leakage, 1041–1042, 1044
Learning, 332, 343, 350, 458–465, 469–473, 477, 478, 486, 853, 857
 curve, 218
 economy, 462
 regions, 381, 457–473, 492, 498
Leisure, 29, 37–41, 63, 110, 112, 113, 149, 686–687, 694, 696, 743
Leontief expansion, 1515
Leontief-inverse, 879, 880, 885–887, 898
Leontief-paradox, 909
Levels of response, 1622
Liberal market economies, 460
Life course, 1324, 1330–1332
Life cycle models, 148, 149
Life expectancy, 1088–1089, 1091, 1093
Likelihood ratio, 1556, 1561
Limited and censored, 1619–1633
Limited and censored dependent variable models, 1619–1633
Limiting distribution, 1602, 1604, 1605, 1611–1614
Limiting factor, 1054
Linear model, 380, 381, 473, 652, 653, 657, 662, 663
Linear model of innovation, 651–652
Linear regression, 1436, 1437, 1443–1446, 1450–1453, 1455, 1456
Linear regression model, 364, 1041, 1347, 1436–1437, 1443, 1444, 1446, 1451–1453, 1455, 1483–1486, 1499, 1503, 1512, 1513, 1532, 1542–1543, 1578, 1615, 1638–1640
Line graph, 1268–1270
Link, 788–791, 793–808
Linkage effects, 869, 870
Linked brushing, 1297, 1302–1306
Linked micromap (LM) plots, 1299–1301
Linked views, 1296, 1302–1305
Link flow, 792, 794–796, 798–802, 804–807
Link-route correspondence, 767
LISA. *See* Local Indicators of Spatial Association (LISA)
Living laboratories, 464, 469
LMAs. *See* Labor market areas (LMAs)
LM plots. *See* Linked micromap (LM) plots
Local buzz, 595–596, 654, 659
Local diffusion, 1539, 1541
Local economic diversification, 622
Local embeddedness, 496
Local Indicators of Spatial Association (LISA), 1176, 1297, 1305–1307
Localization, 458, 462–464, 466, 470–471, 635, 636, 639, 640, 642, 644–646
Localization economies, 403–404, 409, 484
Localized industries, 441–443, 446, 451–452
Local labor markets, 20, 36–37, 39, 45, 47–49, 53–55, 67, 68, 70, 89, 115, 395, 482, 642, 680
Locally weighted regression, 1436, 1440, 1441
Local regression coefficients, 1436, 1441, 1444, 1445
Local spatial spillovers, 364, 1540
Local spillovers, 1561
Local wellbeing, 287
Location, 69, 72
 choice, 83, 96–98, 101–102, 232, 234, 320, 508–510, 519–521, 523–525, 549, 562, 748, 749, 760, 839–841, 1044
 game, 836, 838–840
 theory, 318, 508, 509, 511, 525, 726, 834–836, 839–840, 848, 1283
'Locational hysteresis', 615
Locational privacy, 722
Locational sorting, 1031, 1037
Location-aware technologies (LATs), 706, 711, 712, 716–717, 721, 722
Location decisions of households, 1550
Location equilibrium, 511, 515, 516, 518–522
Location game, 836, 838–840
Location-specific amenities, 5, 11
Location-specific environmental policies, 1004
Lock-in, 593, 594, 596, 599, 603, 604
Lock-in effect, 586
Logistic function, 822–823
Logistic niche, 822, 828
Logit model, 817, 818, 820, 822–823, 862

- Logit route choice, 771–773, 775
 Loose-coupling, 463
 Loss of resilience's, 1097, 1098
 Lot sizes, 24, 25, 29, 30
 Love of variety, 910–912
 Low-probability risks, 995
- M**
- Machine learning methods, 1178, 1181–1183, 1190
MACML. See Maximum approximate composite marginal likelihood (MACML)
 Macro-economic identity, 877, 884
 Malnutrition, 1068
 Malthusian trap's, 1088
 Mantel test, 1328
 Manufacturing sector, 543, 544, 550, 551, 555, 556, 558–560, 563–565
MAR externalities. See Marshall-Arrow-Romer (MAR) externalities
 Marginal
 augmentation, 1584–1587
 damages, 998, 999
 effects, 1536–1538, 1540, 1542
 effects estimation, 1537, 1540
 pollution damages, 998
 productivity, 853, 854, 866–869, 871
 Marginal physical product (or value), 43, 46
 Marked point process, 1278
 Marked spatial point processes, 1329
 Market-based solutions, 1033–1034
 Markets, 477–484, 486
 area, 36, 41–43, 47–55, 510, 833–849
 crowding effect, 221–222, 226–228
 failures, 1090
 penetration, 397, 406, 410
 power, 43–44, 555
 segmentation, 148, 157–160
 shaping, 472
 theory, 35–56
 Markov chain, 310–313, 613–614, 1197, 1205, 1209, 1373–1375, 1379–1381, 1402, 1420, 1448, 1572, 1575–1578, 1580–1584, 1612, 1633, 1643, 1655
 Markov Chain Monte Carlo (MCMC), 1197, 1205, 1206, 1209–1211, 1214, 1215, 1448, 1571–1594
 Markov chain process, 613–614
 Markov chains, 310–313
 Markovian strategy, 960, 961
- Markov-perfect Nash equilibrium (MPNE), 961–963, 965, 966
 Mark-up factor, 44
 Marshall, A., 441–443, 451–453
Marshall-Arrow-Romer (MAR) externalities, 442–443, 446, 449–452
Marshall-Arrow-Romer (MAR) knowledge spillovers, 655, 656
 Marshallian demand, 40
 Marshallian externalities, 491, 496
 Marshallian industrial district, 439–455, 492
 Mass balance principle, 1010, 1012
 Matching benefits of thick markets, 913
 Matching function, 60, 65–70
 Material
 balance, 1011–1012, 1016–1019, 1026
 flow, 1013
 flow analysis, 1012–1015
 Materials balance model, 1009–1026
 Matérn correlation, 1411
 Mathematical programming, 692
 Matlab, 1478, 1487, 1488
MAUP. See Modifiable areal unit problem (MAUP)
 Maximum
 entropy, 890
 principle, 932
Maximum approximate composite marginal likelihood (MACML), 1628, 1630, 1631
Maximum likelihood (ML) estimation, 311, 1512, 1516, 1523–1525, 1527, 1530–1532, 1538, 1554–1568, 1598, 1599, 1602, 1607, 1608, 1611, 1612, 1628, 1658, 1660
Maximum simulated likelihood estimation (MSLE), 1620, 1628–1629, 1631
 May model, 819
 May's logistic Eq., 819
 May's logistic model, 820
MCMC. See Markov Chain Monte Carlo (MCMC)
 MCMC theory, 1575–1579, 1582, 1584
 Mean model, 1549
 Measurement errors, 303, 305, 671, 1464
 Measures, 1254–1258, 1260–1265, 1268, 1270, 1272
 Mechanisms, 477, 480–483, 485, 486
 Mediating variables approach, 1095, 1096, 1098
 Meta-analysis, 304, 989, 990
 Method of moments estimation, 1597–1616
 Methods, 1640, 1643

- Metropolis-Hastings (MH), 1575, 1578–1581, 1583, 1586
sampling, 1423
- MH. *See* Metropolis-Hastings (MH)
- Microblogs, 733
- Micro-data, 7, 9, 13, 14, 664
- Microeconomic data, 436
- Microeconomic theory, 817, 818
- Micromaps, 1299, 1301, 1308
- Microsimulation, 717, 719, 720, 723, 752, 753
- Microsimulation model, 717, 719, 720, 752, 753, 1236–1237, 1239, 1240, 1249–1250
- Migrant-attractive power, 11, 13
- Migrant self selection, 11–12
- Migration, 18–21, 23–24, 27, 37, 53–55, 182–184, 189, 190, 326, 332, 334, 336, 338, 340, 851–872, 1067
cost, 853, 854, 856, 861, 864–866, 871
decision, 549
efficiency, 27
firms, 334, 414
households, 414, 1001
labor, 4
people, 116, 153, 441, 852
- Migration/latency, 1312, 1317, 1324, 1327, 1328, 1332
- Mincer, 52, 321
- Minimum spanning tree, 1261, 1266
- MINITAB, 1478, 1487, 1489
- Minority groups, 96, 98
- Misallocation, 641, 642
- Mitigation, 937
capacities, 1050, 1055–1065
policy, 564, 938
- Mixed index approach, 148, 160
- Mixed index model, 162
- α -Mixing, 1615
processes, 1615
spatial processes, 1599, 1615
- Mixture models, 310
- ML. *See* Maximum likelihood (ML)
- MLP. *See* Multi-level perspective (MLP)
- MMT. *See* Modern monetary theory (MMT)
- Mobile telephony, 709
- Mobility, 96, 97, 103, 104, 706–709, 711–714, 716, 717, 720–723
of capital, 968, 969
cost, 89, 542, 706
histories, 1312, 1330, 1332
of labor, 118, 120, 231, 260, 423
mining, 711, 716, 721
- patterns, 13, 118, 481, 755, 1313
transition model, 1099
of women, 153
- Modal
composite cost, 777
dispersion, 775, 776
- Model-based geostatistics, 1461–1474
- Modeling, 1217–1231
- Model of supply and demand, 110, 112–114
- Models, 812–830
appropriateness, 1214, 1215
extensions, 570, 581–582
selection, 1215
- Modern monetary theory (MMT), 272
- Modifiable areal unit problem (MAUP), 1110, 1113, 1157–1170, 1280, 1299, 1474, 1478
- Modified gravity models, 6, 8, 10
- Module, 1263–1264
- Monocentric cities, 542–543, 552, 554, 566
- Monocentric model, 76–81, 84, 85, 87, 88
- Monopolies, 480
- Monopolistically competitive markets, 216
- Monopolistic competition, 540, 543, 547, 567, 910–913
- Monopoly, 910–912
- Monopsonist, 46
- Monopsony, 46–47, 53
- Monopsony, monopsonist, 46, 47, 53
- Monte Carlo experiment, 1528, 1586
- Moral hazard, 358
- Moran
coefficient, 1176, 1478–1480, 1482, 1483, 1485–1489, 1492, 1494–1497, 1499, 1500, 1502, 1505–1506, 1525
scatterplot, 307, 313, 1305, 1306, 1379, 1480, 1505
- Moran's I, 1176, 1281, 1282, 1305
- Morning peak commuting period, 764, 768, 779, 780
- Mortgage, 89, 97, 103, 119, 134, 136, 139–143
- Motorway, 340
- Movement costs, 1003
- Move-or-stay decision, 152
- Moving average, 1518–1520, 1525
- Moving average process, 306, 1518–1520, 1525, 1541
- Moving-to-Opportunity program, 104
- Moving toward high rent locations, 1003
- MPNE. *See* Markov-perfect Nash equilibrium (MPNE)
- MRIOTs. *See* Multi-regional IO tables (MRIOTs)

- MSLE.** *See* Maximum simulated likelihood estimation (MSLE)
- Multidirectional, 1512, 1525, 1532
- Multifactorial causes of disease, 1315
- Multilayered, 1631, 1633
- Multilayer niche, 822
- Multi-level
- data structures, 1337–1339, 1356
 - methods, 1336–1339, 1348, 1354, 1356–1358, 1360
 - models, 1426
 - regression model, 1345
 - statistical approach, 1352, 1360
 - statistical models, 1345
- Multi-level perspective (MLP), 466
- Multi-market hedonic analyses, 1004
- Multinational corporations, 481
- Multinomial, 1622, 1623, 1625, 1627, 1628, 1632
- Multinomial logit model, 690
- Multiple
- employment centers, 78, 79
 - equilibria, 307
 - membership designs, 1339, 1357
 - possible equilibria, 613
 - testing problem, 1288
- Multiplier, 1515–1516, 1518, 1520, 1525
- Multiplier effect, 678, 1515–1516, 1518
- Multi-regional computable general equilibrium (CGE), 896–897
- Multi-regional IO model, 889, 898
- Multi-regional IO tables (MRIOTs), 888–890
- Multi-site selection models, 982–983
- Multi-stage decision-making, 348
- Multivariate fields, 1390, 1391, 1393
- Multivariate responses, 1338
- Mutation, 458, 468
- Muth condition, 77, 83, 84
- Myers, D.E., 1391
- N**
- NAFTA. *See* North American Free Trade Agreement (NAFTA)
- Nash-bargaining, 66
- Nash equilibrium, 238–240, 242–244, 246, 250, 257, 796, 807, 953, 954, 958, 960–963
- National capitalization rate, 28
- National Center for Geographic Information and Analysis (NCGIA), 1115–1116
- National innovation systems (NIS), 490
- Natural-experimental empirical method, 1030–1031
- Natural experiments, 280, 341, 342
- Natural resource management, 930, 985
- Natural system, 930, 1068, 1074, 1080, 1081, 1220, 1369
- NCGIA. *See* National Center for Geographic Information and Analysis (NCGIA)
- Near epoch dependence, 1615
- NEG. *See* New Economic Geography (NEG)
- Negative equity, 142
- Negative externalities, 204, 1033–1034
- Negative feedback effect, 894
- Neighborhood, 1256, 1258, 1259, 1265, 1267
- change models, 148, 149, 151
 - contextual effects, 1320–1322, 1329, 1330
 - effects, 1342, 1349, 1350, 1356–1358
 - peer effects, 670, 674, 675
 - predictors, 1346
- Neighboring regions, 1536–1539, 1542, 1545, 1551
- Neighbors, 1284, 1285, 1288, 1290, 1292, 1514, 1515, 1517, 1519, 1523
- Neo-classical
- growth model, 292, 293, 298, 299, 313
 - labor supply theory, 60
 - models of growth, 169–190
 - production function, 170–174, 176, 187
 - trade model, 907
- Neogeography, 1108, 1117, 1119–1120
- Neo-Marshallian industrial districts, 440–441, 443–445, 449–455
- Nested
- data structures, 1337, 1348
 - logit, 718
 - logit models, 690
 - sampling, 1391
- Net in-migration, 20, 27
- Net migration, 6–8, 11, 23, 27
- Net operating surplus, 877
- Net wage, 12
- Network Analysis in Geography, 727
- Networked innovation processes, 463
- Networks, 272–274, 393–394, 398, 400–405, 408, 410, 411, 460, 461, 463, 466, 478, 479, 481–486, 592–597, 599, 601–605, 650, 657–659, 661–664, 812–831, 1253–1272
- advantages, 400
 - analysis, 812, 815, 823–830, 1125
 - centrality, 1255, 1262–1264, 1270
 - concentration, 827
 - contagion, 730
 - cost, 762, 773–779
 - dynamic, 594, 605

- effects, 1654, 1656, 1664, 1666–1672
equilibrium, 790, 793, 795, 796, 801–807
paradigm, 491
science, 500, 501
society, 497, 503, 1254
spillovers, 1666, 1668
systems, 788–789, 792, 797–798,
 801, 808
theory, 812–815, 831
time prism, 714–715
topology, 824–827
user-equilibrium, 759–784
- Neural networks, 1126
Neutral models, 1317, 1325–1326, 1332
New Economic Geography (NEG) models,
 354–356, 360, 429–431, 479, 483,
 569–587, 668, 869, 905, 913
New economics of migration, 6
New Growth Theories, 1090
New institutional economics (NIE), 272, 273
Newton-Raphson, 1558
Newton's gravity law, 817
New trade theory, 905, 910–913, 918
New urbanism, 1080
Niche model, 821
NIE. *See* New institutional economics (NIE)
NIS. *See* National innovation systems (NIS)
Non-arbitrage condition, 219, 867
Non-basic sectors, 265
Non-convergence, 20
Non-ergodic stochastic processes, 613
Noninformative prior, 1196, 1200
Nonlinear, 471, 813–814, 819, 827, 830
Nonlinear complementarity theory, 762, 764
Non-linearities, 997, 999
Nonlinear Polya urn process, 614, 615
Non-market goods, 148, 158–159
Non-monotonic response, 942
Non-nested tests, 1529
Non-parametric, 1513, 1514, 1519, 1520,
 1522, 1523
Non-parametric model, 1433, 1520
Nonrenewable, 931–934, 936, 938–940, 945
Non-spatial gravity model, 1662–1668
Nonspatial proximities, 651, 658–664
Nonstationarity, 1158, 1180, 1181, 1191,
 1279, 1436, 1442–1443,
 1448–1449, 1453
Non-tradable
 consumption services, 543, 544, 554, 556
 goods, 23, 30
Non-use benefits, 1006
Normal attractors, 468, 469
- North American Free Trade Agreement
 (NAFTA), 920
Nugget effects, 1389, 1391, 1392, 1395
Null hypothesis, 1320–1322, 1325, 1329, 1332
Number of children, 25
- O**
- Objective function, 1393–1395, 1397
Objects, 1278, 1284, 1290
Observational equivalence, 304
Observed information matrix, 1557, 1560, 1561
Occupational analysis, 322
Occupations, 318–325
OIE. *See* Old institutional economics (OIE)
Old institutional economics (OIE), 262,
 272, 273
Oligopoly model, 238, 241, 243, 245, 246, 257
Oligopsony, 53
OLNE. *See* Open-loop Nash equilibrium
 (OLNE)
Omitted variables, 157, 160–162, 303, 305,
 1515, 1520, 1528
“One-size-fits-all,” 366–367
One-step GMM estimation, 1599, 1601–1604,
 1612
On-the-job search, 63, 64
Open evaluation, 679–681
Open-loop Nash equilibrium (OLNE), 961,
 963, 966
Open-loop strategy, 960, 961
Opportunity cost, 7, 39, 62, 80, 120, 219, 320,
 508, 544, 547, 687, 932, 980, 985,
 1091
Optimal
 control, 248, 249
 policy, 931–933, 936–941, 943, 944
 size, 204
Optimization methods, 717–719
Optimization procedure, 257, 717–718, 1558
Option value of waiting, 872
Orchestration, 461, 464
Organization, 592, 593, 596, 599, 600, 602, 605
Organizational change, 196, 205
Organizational proximity, 498–500, 658,
 661–663
Organizational science, 194
Organized complexity, 814–815, 828
Origin and destination dependence, 1655, 1659,
 1660, 1666, 1670
Origin-destination (O-D)
 dispersion, 770, 775–778
 flow data, 1297, 1298, 1303

- Origin-destination (O-D) (*cont.*)
 flow models, 152, 1653–1672
 pair, 790, 792–800, 805, 806
- Origin effects, 1664–1671
- Oswald's hypothesis, 119
- Oswald's thesis, 76, 89
- Outbreaks/spread of infection, 1313, 1317, 1322–1323, 1326, 1327
- Outcome path dependence, 611, 618
- Out-migration, 20, 27
- Output
 coefficients, 898
 multipliers, 880
 price multipliers, 882
- Outright owners, 76, 89
- Outsourcing, 913
- Overdispersion, 1292
- Overview + detail, 1149
- Owner occupation, 137, 139–142
- P**
- Page ranking, 735
- PAJEK, 728
- Panel data, 292, 299, 304, 1638, 1641–1645, 1648–1650
- Panel data model, 1178, 1180, 1190, 1572, 1587, 1589, 1643, 1644, 1648, 1649
- Parallel coordinate plots, 1142–1143, 1149
- Parameter
 expansion, 1584, 1585
 heterogeneity, 307–308
 instability, 1522–1523, 1530–1532
- Pareto distribution, 729
- Pareto-optimal, 871
- Partial agglomeration, 554, 560
- Partial linearization, 780
- Passive use values, 976–977, 987
- Patent citations, 481, 482
- Patent data, 376–377, 380, 419, 662, 664, 1090
- Patents, 376–381, 383–384
- Path, 792–795, 798–802, 804–806
 creation, 462, 467, 620, 624, 625
 dependence, 201, 458–459, 466–468, 471, 472, 593, 594, 596–599, 602, 609–627
 dependencies, 912
 flow, 792, 794, 795, 797, 798, 800–806
 interdependence, 458, 462, 466–468, 471
- Pathogen
 ecology, 1318
 strains, 1317, 1318, 1326, 1327
- Patterns of health events, 1312–1317, 1323–1328, 1332
- Payment for environmental services programs, 1040–1041
- Payoff matrix, 343–346
- PCA. *See* Principal components analysis (PCA)
- Pecuniary external effects, 480, 483–484
- Penalized regression, 1455
- Penalized splines, 1409
- Percentage owner-occupied, 28
- Perceptions, 996, 999, 1000, 1007
- Perceptions of environmental benefits, 1007
- Perfect competition, 853, 854, 857, 860
- Periodicities, 1387, 1397
- Peri-urban, 1101
- Permafrost, 1058
- “Permissible,” 1464, 1465
- Personal unemployment, 9
- Pervasive computing, 709
- Phase diagram, 174, 188
- Physical distance, 651, 653–654, 660
- Physical geography, 1055
- Physician behaviors, 1315, 1322
- Pipelines to distant knowledge, 380
- PIT. *See* Probability integral transform (PIT)
- Pivots, 1566
- Place-based, 1548
- Place-based policies, 104, 106, 111, 121, 122
- “Place-tailored” policies, 369
- Planar, 1257–1260, 1266, 1268, 1270–1271
- Platform, 459, 460, 463–465
- Platform model of regional innovation policy, 463
- Platform policies, 657
- Point data, 1197–1202, 1206–1213
- Point (pattern) data, 1197–1202, 1427
- Point pattern, 1278, 1427
- Point process, 1412–1413, 1415
- Point process models, 1412–1413
- Poisson
 distribution, 63, 825, 1198, 1199, 1202, 1205
 kriging, 1290, 1291, 1469, 1471, 1473
 regression, 1500
- Policy, 415, 416, 423, 432, 435, 436, 1236–1240, 1245, 1250, 1251
 consequences, 570, 586–587
 constraints, 119–121
 emergence, 470–472
 evaluation, 1038, 1045
- Pollution, 565, 566
 control, 936, 939
 damage, 963, 996–1000
- Polycentric city, 543, 552–554, 566

- Polygon data, 1198, 1212
Poor neighborhoods, 94, 105
Population, 332–343
 bottleneck, 1317, 1318, 1327
 density, 544, 545, 547, 550, 555, 560,
 564–566
 growth, 1086–1095, 1100–1102
 growth rates, 156
 migration, 54, 1279, 1505
Porterian diamond, 451
Positional error, 1323–1324
Possible wage and rent combinations, 1002
Post-epidemiology, 1312–1313
Posterior
 distribution, 1196, 1199–1202, 1204, 1205,
 1209–1214, 1632, 1633
 residuals, 1347, 1348
 standard deviation, 1206, 1207, 1212
Potential destinations, 18, 22
Poverty level, 1343
Power-form, 825, 826, 829–830
Power law, 729
The Power model, 1466
Preadaptation, 469
Precipitation events, 1052
Precision matrix, 1563
Precision-weighted estimation, 1349
Prediction
 error, 1439
 uncertainty, 1390
Predictive integral transform, 1409
Pre-epidemiology, 1312–1313
Preferences, 478, 485
Preferential attachment, 729, 826, 1267–1268,
 1270
Preservation/existence values, 32
Prey–predator, 821
Price elasticity, 267, 693–696, 1648
Price-quantity interaction, 896
Primal-dual, 882
Primary input multipliers, 880, 882
Primate city, 646, 647
Principal-agent issue, 701
Principal components analysis (PCA), 1183
Prior, 1631–1633
Prior distributions, 1196, 1199, 1201, 1202,
 1204, 1209, 1213, 1402, 1406,
 1407, 1410
Prior probability, 1199, 1201, 1202, 1209
Private good, 999
Probability integral transform (PIT), 1407,
 1408, 1415
Producers of housing, 156
- Product
 cycle, 397, 404, 410, 643
 innovation, 396–399, 403, 409–410, 426,
 638, 639, 642
 varieties, 216, 224, 234, 392–394, 398, 399,
 403, 406, 410
Production, 333–342, 345, 349, 350,
 477–483, 486
 chains, 913
 factor, 477
 factor mobility, 908
 function, 334, 340, 355, 359, 365, 366
 networks and strategic alliances, 493
 possibility frontiers, 905–909
Productivity, 333–334, 336, 340, 342, 345,
 350, 476, 477, 484, 632–639,
 641–644, 646
 changes, 114
 decomposition, 642
 of workers, 99, 104
Product-life cycle, 598, 650, 655, 656
Product vintage, 392
Profile log-likelihood, 1562, 1563
Property taxes, 150, 158
Property value or rent compensation, 998
Proportionality, 760, 770, 781–782
Proximity, 394, 400–402, 408, 490–492,
 496–501
Proximity advantage, 400
Public
 expenditure, 12, 259, 269–272, 275, 359
 goods, 483, 651–653, 974–977, 983–985,
 987–991, 998–999
 health data, 1313–1314
 policy, 31
 services, 148–150, 158, 326, 354, 1639
 spending, 355–359, 362–366, 369
 transport, 96, 98, 99, 102–105, 760,
 762–764, 769, 773–775, 778, 780
 transportation, 690, 693–697
‘Punctuated equilibria’, 615, 617, 624
Purchases structures, 877, 884
Pure science, 1109, 1110
Purposive actions, 620, 623, 625
- Q**
Qualitative approaches, 273
Quality of Life, 1002
Quantile, 1140, 1141, 1147
Quantitative Revolution, 1280–1284
Quasi-dynamic model, 751–752
Quasi-empirical method, 1031

- Quasi-experimental, 1039, 1040, 1045
 Quasi-experimental method, 1039–1040, 1045
 Quasi-maximum likelihood, 1562
 Quasi-public good, 651, 653
- R**
 R, 1442, 1446–1448, 1450, 1451, 1453,
 1454, 1456
 Race, 94, 95, 99
 Race to the bottom, 1002
 Ramsey-Cass-Koopmans model, 170, 185–190
 Ramsey's formula, 944
 Random
 coefficient models, 1351
 effects, 1447
 graph, 1265
 matching, 78, 81, 86
 network, 813–814, 820, 824–828
 sampling, 1386–1388
 utility maximization, 818
 variables, 1346, 1348, 1352, 1462
 Random function (RF) model, 1462, 1463,
 1470, 1473
 Random-intercepts, 1346–1348, 1351
 Randomization, 673, 674, 677, 680–681
 Randomness, 1325
 Random-slopes, 1351, 1352
 Range, 1465, 1466, 1470, 1473
 Rank, 817, 818, 828
 Rank-size model, 829, 830
 Rank-size rule, 729, 815–818, 827, 829
 Rare earth metals, 1010, 1019–1025
 Rare earths, 1023–1024
 Rate of convergence, 297–308
 R&D activity, 199, 339, 382, 395–396,
 404–405, 417, 421, 1591
 R&D co-operation networks, 493
 Real estate developers, 128–129
 Real estate investment, 127, 134
 Real estate market efficiency, 128
 Reallocation, 638, 641, 642
 Real option, 332, 342–350
 Recall, 61, 63, 64
 Reciprocity, 493
 ‘Recombinant’ path dependence model, 619
 Recombinations, 459, 460, 463–465, 470
 Rectangular, 1387
 Recurrent events, 937
 Recursive cumulative causation models,
 612, 613
 Recursive modelling, 688
 Redistributive income growth, 894
- Reference group, 285–287
 Reframing, 467
 Regime shift, 933, 937, 938, 942, 945
 Region, 540–545, 549, 551, 561, 566
 Regional, 478, 481–483, 485–486
 advantages, 461
 agglomeration, 461
 authorities, 358
 balance of payments, 267
 control of pollution, 967–968
 convergence, 278, 292, 301,
 1367, 1374
 data, 1290
 development, 446, 452, 1374–1375
 economic development, 413–436
 experimentalism, 463
 experiments, 463
 growth, 277–288, 332–334, 337–342,
 349–351
 growth model, 1573
 inequalities, 354, 355
 innovation policies, 385–387, 654
 innovation strategy, 462
 institutional differences, 209
 labor market, 7, 36, 37, 42, 45, 47–49, 52,
 56, 121, 880, 1163
 milieus, 392, 402–405
 path dependence, 596–599, 612, 616
 platforms, 463–465
 policies, 354–361, 365–367, 369, 668
 political science, 698
 resources, 260
 spatial polices, 121
 specialization, 111, 270, 595,
 609, 905
 unemployment, 109–123
 Regional innovation systems (RISs), 381–382,
 387, 457–473, 490–491, 493
 Regional input-output table, 876, 877
 Regional purchase coefficient (RPC), 888
 Regional systems of innovation (RSI), 651,
 656, 657, 659–661
 Regression
 coefficients, 1436–1438, 1441–1449, 1451,
 1453, 1455–1457
 discontinuity designs, 677
 models, 1282–1284, 1286, 1288, 1289
 Regularization, 1465, 1470–1471
 Regular lattice, 1198, 1212
 Relatedness, 458, 459, 462, 464, 465, 467,
 468, 470
 Related variety, 379, 386, 458, 464, 467–469,
 656, 657, 659

- Relational
 economic geography, 591–605
 proximities, 498, 500
 space, 498, 499
- Relational-evolutionary perspective, 592, 594, 604, 605
- Relationship, 876–882, 891, 892
- Relative risk, 1425, 1429, 1431
- Relative unemployment discrepancy, 7
- Relativities, 278–280, 282–287
- Relocation of firms, 97
- Remittances, 872, 1099, 1100
- Renewable resource, 933, 934, 938, 945
- Renewal activities, 397–398, 409
- Rents, 19, 21–30, 32
- Rent seeking, 358
- Repeated measures, 1338, 1358
- Repeat sales method, 159
- Replacement level, 1091, 1093
- Representative fraction, 1113
- Research and development (R&D), 376–386, 388, 393–401, 405, 407–409, 411, 477, 486
 intensity, 395, 398–400, 409
- Research collaboration, 481, 482
- Reservation wage, 62–64, 69–72
- Residential
 histories, 1314, 1330–1331
 mobility, 88, 89, 96, 97, 104, 147–163, 1330
 segregation, 94, 99, 105–106
- Resilience, 625–626, 812, 827–829
- Resilient networks, 829
- Response to health care policy, 1322
- Retired, 1005, 1006
- Returns knowledge, 477
- Returns to education, 26, 36, 43, 52, 637
- Return to R&D, 397–398, 409
- Return to unionization, 26
- Revealed preference, 688, 690–692
- Revealed preference method, 975–977, 980, 987–991
- Revealed relatedness, 471
- Reverse nested combined model, 778
- Reweighting, 1236, 1240–1241, 1243–1247, 1250, 1251
- RF model. *See* Random function (RF) model
- Ridge regression, 1442, 1446–1447
- R-INLA, 1414, 1415
- Ripple effects, 1065
- Rising income nationwide, 23
- Risk averse, 89
- Risks of damage, 31
- RISs. *See* Regional innovation systems (RISs)
- Road, 1268–1271
- Road network, 760, 761, 764–779, 781, 782
- Robust, 999
- Robust LM test, 1532
- Robustness, 303–304
- Robust tests, 1512, 1524, 1526, 1528, 1532
- Root mean squared, 1439, 1451, 1452, 1456
- Roughness penalty, 1204
- 1st round indirect effects, 879
- Route factor, 1260, 1261
- Routines, 392, 393, 396, 398, 404, 406, 408
- RPC. *See* Regional purchase coefficient (RPC)
- R programming language, 1408, 1413, 1415
- RSI. *See* Regional systems of innovation (RSI)
- Runoff, 1052, 1054
- Rural region, 855, 856
- Rural-urban migration, 4, 5, 14, 856, 857, 860, 872
- S**
- Saddle path, 866, 868
- Sahel desert, 1063
- Sales structures, 877, 884
- Sample density, 1388
- Sampling, 1278, 1279, 1281, 1290, 1291
 clustering, 1387
 configurations, 1391, 1392
 coverage, 1387
 densities, 1392
 priorities, 1395, 1397
 variance, 1387
- Sanitation, 1066–1068
- Santiago, 762, 780, 784
- SAS, 1478, 1488, 1491
- Savings rate, 173, 175, 184, 187–190
- Scale, 1107–1108, 1113–1114, 1117
- Scale effect, 172, 199, 268, 414, 530, 639, 1157–1168, 1537
- Scale-free Internet myth, 729
- Scale-free networks, 824–829
- Scatter plots, 1139–1143, 1149–1150, 1443, 1444, 1453, 1454
- Scenario analysis, 350
- SCGE. *See* Spatial computable general equilibrium (SCGE)
- Scientific inference, 1314–1315, 1325, 1333
- Scientific visualization, 1287
- Score, 1557, 1559, 1560
- Scrolling or panning, 1149
- SDBMS. *See* Spatial database management systems (SDBMS)

- Search, 76, 79–82, 85, 86, 88–89, 458, 472
 intensity, 63, 64
 models, 152
- Secondary
 employment districts, 552
 products, 890
 variables, 1393, 1394
- Second demographic transition, 1091, 1092, 1099
- Second Law of Geography, 737–738
- Second-order neighboring observations, 1539, 1542, 1546
- Second-phase, 1393–1395, 1397
- Second-phase spatial, 1393
- Sectoral specialization, 905, 910
- Sectoral systems of innovation, 656–657
- Segmentation, 286
- Selection, 467, 468, 473
- Selection and retention, 458
- Selective migration, 5
- Selectivity bias, 27
- Self-organization, 458, 472–473, 812, 824, 825, 828
- Self-organizing maps (SOM), 1183, 1190
- Self-reinforcing effects, 1537
- Self-reinforcing ‘lock-in’, 610, 612–618, 621
- Self-selecting migration, 116
- Self-sufficiency ratio, 888, 893
- Semi-arid, 1060
- Semivariance, 1463–1465
- Semivariogram, 1388–1390, 1395, 1463
- Sensemaking, 469
- Separable, 768–769, 784, 1561
- Sequential travel forecasting procedures, 784
- Shadow cost, 817
- Shadow price, 39, 186, 932
- Shock propagation, 830
- Shrinkage, 1446–1447
- Shrinkage estimates, 1349
- SI. *See* Systems of innovation (SI)
- Sickle cell trait, 1317–1318
- Signed root deviance, 1557, 1561
- “Sill” parameter, 1465
- SIM. *See* Spatial interaction models (SIM)
- Simulated annealing, 1236, 1240, 1243–1245
- Simulation, 1218, 1222, 1225–1227, 1229–1231
- Simultaneity, 294, 302, 303
- Simultaneous spatial autoregressive (SAR) model, 1212, 1280–1281, 1419
- Singular value decomposition, 1444, 1445
- Site selection, 1032, 1045
- Size, 1386–1388, 1392–1394
- Skills, 110–111, 115–117, 121, 122, 317–328, 476, 478, 481, 482, 650, 663, 664
- Small-area, 1236, 1238, 1246, 1250
- Small island nations, 1060, 1064, 1069
- Small numbers problem, 1290, 1318–1320, 1325
- Small-world, 1267, 1268, 1270
 networks, 824
 phenomenon, 729
- The Small world problem, 728
- Smoothing, 1436, 1442, 1452–1453, 1456
- S-O. *See* System-optimization (S-O)
- Social, 1254, 1258, 1263, 1264, 1267, 1271, 1272
 capital, 460, 462, 463
 and cultural distance, 479
 distance, 730
 equity, 747, 1076
 filters, 661
 geography, 745, 755
 indifference curves, 905, 907, 908, 917, 918
 media, 711, 721
 network analysis, 502, 725–739, 752, 823–826, 1263, 1264
 networks, 483, 662, 707, 708, 710, 717, 720–721, 723, 725–739, 823–826
 norms, 1098, 1100
 proximity, 658, 661, 662
- Sociocultural and institutional assets, 442
- Socio-cultural regime, 464, 466
- Sociograms, 728
- Socio-institutional and geographical conditions, 650
- Soft innovation, 378
- Software package, 238, 727, 728, 894, 900, 1133, 1139, 1299, 1306–1308, 1412, 1486
- Soil fertility, 994
- Solow-Swan model, 170–185, 188, 190
- SOM. *See* Self-organizing maps (SOM)
- Sorting, 486, 634, 635, 637, 640
- Sources of spatial autocorrelation, 1312, 1315–1325
- Space, 1029–1039, 1041–1046
- Space-time, 1365–1382
 cube, 1152–1154, 1175, 1188
 data analysis, 1365–1382, 1429–1431, 1641
 geostatistics, 1178, 1181, 1190
 interpolation, 1181
 model, 1124, 1176–1178, 1181, 1183, 1190, 1429–1431, 1433, 1512, 1591, 1642
 path, 711–716, 720, 722
 prism, 711, 713, 714, 716, 722
 series, 1174, 1177, 1184
 visualization, 1175, 1176, 1185–1191
- Space-time analysis of regional systems (STARS) platform, 1175

- Space-time autocorrelation function
(ST-ACF), 1177
- Space-time autoregressive integrated moving average (STARIMA), 1177–1180, 1190
- Space-time geographically weighted regression (GWR), 1180, 1190
- Sparse, 1555, 1562–1564, 1566–1568
- Spatial, 1620, 1621, 1624–1633
analysis, 1159, 1163–1165, 1169, 1170
autocorrelation, 301, 303, 312, 313, 341, 361, 362, 654, 737, 1035, 1134, 1157–1158, 1162, 1165–1167, 1170, 1279, 1281, 1282, 1286, 1289, 1292, 1297, 1305, 1315–1325, 1327, 1358, 1375, 1377, 1379, 1410, 1437, 1438, 1448, 1449, 1452, 1477–1506, 1512–1514, 1517–1528, 1530–1532, 1624, 1630, 1640
autocorrelation in the residuals, 1327
autoregressive, 1627, 1628
autoregressive local estimation, 1523
autoregressive process, 1180, 1517–1518, 1521, 1541, 1665, 1668
clustering, 1420, 1421, 1433
competition, 567
concentration, 476
consumer surplus, 25, 26, 1006
correlation, 1197, 1198, 1207–1208, 1214, 1420, 1421, 1425, 1427, 1430, 1432
covariance, 1427, 1428, 1462, 1466
data, 669–673
data mining, 1280, 1289–1290
data types, 1197–1198
dependence, 1111, 1389, 1390
differencing, 678
diffusion, 402, 653, 654, 752, 1283, 1514, 1515, 1518, 1541, 1587
disparities, 669, 670, 672–674, 680, 681
distortions, 121
disturbance process, 303, 306
dynamic panel data, 1587
dynamics, 1365–1382
econometric approach, 357, 358, 367
econometrics, 148, 161, 162, 301, 302, 321, 360, 363, 364, 366, 367, 500, 670, 672, 673, 1030, 1046, 1180, 1277, 1280, 1282–1286, 1306, 1367, 1419–1423, 1429, 1512, 1516, 1528, 1532, 1535–1551, 1568, 1572–1573, 1590, 1598, 1638, 1653–1672
economics, 812, 815–821, 823, 824, 826–830
effects, 1512–1523, 1532
embedding, 1257–1260, 1271
epidemiology, 1420, 1424–1427, 1429
equilibrium, 17–33, 332–340, 342, 349, 350
equilibrium models, 789–790
error model, 1421, 1423, 1537–1539, 1544, 1546–1548, 1554, 1559, 1560, 1562, 1563, 1565
filtering, 1477–1506, 1624–1627, 1633
frictions, 566, 567
gravity model, 1664, 1667–1671
heterogeneity, 1029–1034, 1041–1044, 1111, 1180, 1279–1280, 1284, 1286–1287, 1325, 1371, 1426, 1456, 1511–1513, 1522–1524, 1530–1532, 1574
interaction, 816–819, 1500–1502, 1505
interpolation, 1111, 1461–1474
kernel, 1428, 1432
Keynesianism, 272
labor market equilibrium, 17–33
lag model, 301–303, 362, 363, 1180, 1491, 1496, 1514–1517, 1527–1531, 1643–1644
lags, 1540–1544, 1598–1600, 1606–1615
microsimulation, 752, 753, 1235–1251
mismatch, 37, 70, 82, 93–106
mixing, 1599, 1615
models, 816, 1401–1415
moving average, 1537–1539, 1541
network analysis, 1253–1272
network models, 1264–1270
networks, 811–831
nonstationarity, 1436, 1453
oligopoly, 238, 241–243, 257
panel data models, 1178, 1180, 1190
panel models, 1637–1651
panels, 1637–1651
pattern, 1441, 1452–1453
point patterns, 1297, 1298, 1307
Poisson process, 1198
policy, 668–669, 671, 674, 676, 681
probit model, 1423
process, 1436, 1447, 1448, 1458
production cycles, 913, 914, 918–920
proximity, 419, 424
regimes, 1288–1289
regression models, 1282–1284, 1286, 1288, 1289, 1511–1532, 1535–1551, 1561, 1626–1627, 1633
resolution, 1113, 1120
sampling, 1385–1397
shock, 1030, 1031

- Spatial (*cont.*)
 sorting models, 1042–1043
 specific effects, 1642
 spillover effects, 1638, 1645–1646,
 1648, 1649
 spillovers, 268, 272
 statistics, 161–163
 structure, 1391, 1393, 1395
 theory of unemployment, 95, 105
 variation, 1388, 1393
 weight matrix, 301, 304, 364, 1177, 1179,
 1481–1483, 1486–1491, 1494, 1496,
 1506, 1513, 1514, 1521, 1524, 1539,
 1543, 1547, 1548, 1563, 1564, 1574,
 1586, 1588, 1626, 1660, 1662–1663,
 1668
 weight matrix eigenfunctions, 1481
- Spatial-autoregressive autoregressive (SARAR)
 (1,0) model, 1611, 1612
 (1,1) model, 1607, 1612
 (p, q) model, 1612
- Spatial autoregressive (SAR) model,
 1212–1214, 1421–1423, 1429
 parameters, 1604, 1607–1609
 regression, 1420–1423
- Spatial clustering of
 consumption levels, 286
 innovative activities, 653
- Spatial computable general equilibrium (SCGE), 260
- Spatial concentration of
 firms, 442
 production, 509, 520
- Spatial database management systems (SDBMS), 711
- Spatial Durbin error model, 157, 161, 162,
 1518, 1520, 1526, 1529, 1561, 1562,
 1573, 1574, 1584–1587
- Spatial Durbin model (SDM), 157, 161, 162,
 302, 303, 363, 364, 1518, 1520,
 1526, 1529, 1549–1551, 1561, 1573,
 1623, 1641, 1643, 1646–1650
- Spatial interaction
 data analysis, 1500–1502
 effects, 292, 1638, 1639, 1648
 models, 152, 744, 746, 747, 815, 816, 1126,
 1654–1658, 1662–1671
- Spatial-interaction location models, 744,
 746–748, 750–752
- Spatial interaction models (SIM), 152, 744,
 746, 747, 815–820, 822–823,
 828–830
- Spatially adaptive kernels, 1320
- Spatially varying coefficient (SVC) model,
 1289, 1447, 1448, 1458
- Spatially varying relationships, 1436, 1457, 1458
- Spatial-temporal outlier, 1184
- Spatio-temporal, 1138, 1150, 1152–1153,
 1284, 1293
 analysis, 1126
 association rule mining, 1174
 clustering, 1174, 1175, 1184, 1190
 data, 1297–1298, 1307
 data mining, 1173–1192
 model, 1572, 1587–1593
 sequential pattern mining, 1174, 1175
- Spatio-temporal data-mining (STDM),
 1173–1192
- Spatio-temporal forecasting, 1174
- Spatio-temporal scan statistics (STSS),
 1184–1185, 1190
- Spatio-temporal visualization, 1174
- SPDE. *See* Stochastic partial differential equations (SPDE)
- spdep* package, 1486
- Specialization of employment, 85
- Specializations, 479, 482, 484–486, 651, 652,
 655–663
- Speciation, 468
- Speciation and learning, 458
- Specification search, 1528, 1529
- Specific ecologic effect, 1340
- Specific-to-general, 1528, 1529
- Speculative migration, 69, 70
- Spherical model, 1464, 1465
- Spillovers, 304–342, 356, 357, 360, 362, 364,
 366, 479, 481, 483–485, 678, 1030,
 1035, 1040–1046
- Spillovers and feedback, 888
- Spin-off, 481–483, 614, 616, 621
- Splines, 1409, 1413
- Sprawl, 1095, 1100
- Spreading, 1387, 1391, 1392
- S-shaped, 407, 410
- Stable equilibrium, 261, 262
- ST-ACF. *See* Space-time autocorrelation function (ST-ACF)
- Stackelberg equilibrium, 957, 959
- Stage at diagnosis, 1312, 1321, 1326
- Stakeholders, 367–369, 464, 535, 788, 1135
- Standard errors, 1440, 1442, 1451, 1452, 1456
- STARIMA. *See* Space-time autoregressive integrated moving average (STARIMA)
- Stated preference, 690–692

- Stated preference methods, 690, 975, 977, 983, 987, 988, 990, 991
Stated preference valuation, 1037–1038
State variables, 300, 302
Static complexity, 813, 815–819
Static microsimulation, 1237–1239
Static network oligopoly, 244–245
Stationarity, 306, 311
Stationary, 1462, 1463
Stationary process, 1207
Statistical graphics, 1139–1143, 1147–1150
Statistical inference, 1436, 1442–1443, 1447, 1458
Statistical nonstationarity, 1158
Steady state, 170, 174–179, 181–184, 188–190, 292–298, 300–303, 307, 310, 864–866, 933–935, 938–939, 942
Steady-state growth rate, 262, 266–267
Stiglitz Report, 278
Stochastic
 dynamics, 929–946
 effect, 226
 kernel, 311–312
 process, 1513
 route choice, 761, 769–773, 775, 778, 781
 variation, 753
Stochasticity, 257
Stochastic mode and deterministic route choice, 772, 775
Stochastic partial differential equations (SPDE), 1411
Stochastic user-equilibrium (SUE), 770, 771, 773
Stock pollutants, 963–968
Strange attractors, 468–470
Strategic behavior, 1031, 1042, 1044
Strategy, 396, 398, 399, 407, 410–411
Stratified random, 1388
Stratified sampling, 1386, 1388
Streamflows, 1061
The Strength of weak ties, 728
Strong inference, 1312, 1315, 1325, 1332
Structural
 equivalence, 1264
 funds, 356–358, 363, 366, 367
 holes, 463, 464
Structuration perspective, 601
Structuration theory, 597
Structure and relations between people, 1548
STSS. *See* Spatio-temporal scan statistics (STSS)
Study designs, 1343, 1358
Stylized models of local economic development, 455
Subgraph, 1263–1264
Subnational level, 354, 356, 359, 361
Subsidence, 1057
Subsidies, 697, 698
Substitution effect, 40, 189, 358, 558
Substitution of capital for land, 29
Suburban, 476, 1050
Suburbanization, 94, 97
SUE. *See* Stochastic user-equilibrium (SUE)
Supernetwork, 790, 794, 796–797, 804–805
Supplier networks, 493
Supplies of clean locations, 1006
Supply
 chain equilibrium, 790, 797, 801, 808
 chain network, 787–809
 chains, 696, 787–809, 880–881, 897, 920
 decomposition, 40
 and demand pools, 890
 elasticity, 43, 47, 158
 of labor, 20–22, 26, 29, 1001, 1004
Supply and use table (SUT), 890–892
Supply-side growth models, 260
Supply-side influences, 23
Support vector machine (SVM), 1174, 1181–1183, 1190
Sustainability, 262, 272, 471, 478, 598, 709, 952, 958, 1053, 1071–1081, 1097, 1115
Sustainable city, 1072–1077, 1079, 1081
SUT. *See* Supply and use table (SUT)
SVC. *See* Spatially varying coefficient (SVC)
SVM. *See* Support vector machine (SVM)
Symbiosis system, 821
Symmetric, 769
Syndromic surveillance, 1313, 1327
Synergisms, 999
Synergistic interactions, 997
Synthetic populations, 1129, 1236, 1237, 1241–1245, 1251
Systematic
 designs, 1387, 1388, 1391
 sampling, 1386–1388, 1397
 unaligned, 1388
System-optimality conditions, 798–799, 802
System-optimization (S-O), 790–793, 796–799, 801–803, 808
Systems integration, 379
Systems of innovation (SI), 209, 214, 381, 457–473, 490, 491, 502, 651, 656, 657, 659, 660

T

- Tacit, 490, 496
 Tacit knowledge, 221, 377, 378, 477–479, 482, 492, 495–498, 502, 653
 Tax, 1236, 1238–1240, 1245
 Taxation, 30
 Tax competition, 968, 969
 Taylor-expansion, 879, 887
 Technical coefficients, 878, 888, 889, 897
 Technological
 change, 195, 196, 199, 205, 649–664
 co-operations, 493
 diffusion, 397–398, 406–409
 learning, 462
 spillovers, 181, 190
 Technology, 18, 22, 31
 Technology spillovers, 181, 408–409
 Telepresence, 707
 Teleworking, 686
 Temporal autocorrelation, 1429
 Temporal relationship, 11
 Tenure choices, 118
 Territorial production system (TPS), 446–447, 450
 Territorial turn, 445
 The Amazon, 1061
 The Intergovernmental Panel on Climate Change (IPCC), 1050–1051, 1055, 1056, 1058, 1060–1062, 1069
 Theory of movements, 862
 Thermodynamics, 1010, 1011, 1024–1026
 Third Italy, 443, 444, 450–453
 Thirlwall's law, 268
 Threshold effects, 1087
 Tiebout-sorting, 149, 155, 162, 337
 Tile, 1112
 Time, 1386, 1387, 1395, 1397
 Time geography, 707–708, 711, 720
 Time-period specific effects, 1642
 Time perspective of interest, 23
 Time series analysis, 1174, 1176, 1178, 1182
 Time-series studies, 696
 Tobin's q, 867, 868
 Tobler's First Law of Geography, 1108, 1109, 1111, 1112
 Tolerance, 327
 Topological network, 824–827
 Total factor productivity (TFP) growth, 397, 408, 409
 Tour, 707–710, 718
 Tour-based, 763, 784
 TPS. *See* Territorial production system (TPS)
 Tradable manufactured goods, 544

- Trade, 1050, 1065, 1066
 coefficients, 888, 890, 897, 898
 cost, 228, 231, 234, 235, 355, 476, 530–531, 583–584
 overlap, 914–918
 triangles, 906, 908
 in value-added, 923
 Traded interdependencies, 492
 Traffic assignment with proportionality, 781
 Traffic equilibrium, 761
 Traffic flow, 699–701, 743, 765, 1174, 1182, 1239
 Trajectory reconstruction, 716
 Transaction costs, 87
 Transboundary externalities, 1092, 1101
 Transboundary pollution, 952, 967, 968
 Transfer, 477–480, 483, 486
 Transition probability matrix, 311–313
 Transition region, 466, 467
 Transparency, 365, 369, 679–680
 Transport, 741–756, 1254, 1255, 1258, 1264, 1267, 1270–1271
 costs, 335, 340, 341, 479, 540–545, 548, 550–551, 869–871
 modes, 83–85
 planning, 752, 755
 policies, 105
 Transportation, 788–794, 797, 799, 804–805, 807–808
 cost, 906, 907, 909, 913, 915
 data, 687
 forecasts, 699
 and land-use plans, 688, 689, 692, 698, 699
 network, 238, 244, 245, 257, 787–809
 design, 784
 equilibrium, 761, 782, 788, 790, 792, 795–797, 801, 804, 808
 model, 793, 808
 systems or modes, 763, 764, 778
 user optimized, network equilibrium, 795
 Transport interaction model, 741–756
 Transversality, 458, 462–464, 470, 471
 Travel
 behaviour, 685–702
 cost approach, 1031, 1037
 demand, 685–702, 761–763, 783
 forecasts, 761, 763, 779, 784
 Travel-cost analysis, 1037
 Travel-cost model, 977, 980–983, 987
 Treatment effects, 672, 676, 678–679
 Tree, 1261, 1265
 Triangular, 1387, 1391–1393

- Trip, 706–708, 710
 distributions, 688
 frequency models, 980–983
- Trip-based, 763, 764, 783
- Trip-based approaches, 706–708, 710, 716, 764
- Tripolar analytical framework, 600
- Tripolar framework of cluster evolution, 604
- Tri-polar framework of cluster evolution, 600, 604, 605
- Truncation/censoring, 1620, 1621, 1623–1624, 1633
- Trust, 127, 134, 318, 401, 419, 421, 434, 449, 451, 453, 490, 493, 494, 502, 653, 1357, 1548
- Turnover rates of the housing stock, 152
- Two-region IO model, 886
- Two-step GMM estimation, 1599, 1601, 1602, 1604–1606, 1609–1612
- Type II, 887, 891, 893, 894
- Type III demo-economic IO model, 894
- Type II interregional IO model, 891–893
- Type II interregional Leontief-inverse, 893
- Type II price model, 893
- U**
- UCINET, 728
- UE. *See* User equilibrium (UE)
- UGC. *See* User-generated content (UGC)
- UI. *See* Unemployment insurance (UI)
- Unambiguously higher wages, 22
- Unambiguously lower wages, 22
- Unbiasedness, 163, 1478, 1490, 1491, 1494, 1495, 1503
- Uncertainty, 332, 342, 343, 348, 350, 872
- Underlying preferences, 31
- Undervalue the amenity, 30–31
- Undesirable areas, 24
- Unemployment, 36, 37, 41–43, 45, 54, 55, 60, 61, 63–70, 855–857, 860, 862
 benefits, 894
 disparities, 110–112, 114, 119, 122
 rates, 4–10, 110–117, 119–123
- Unemployment insurance (UI), 111, 119–121
- Unemployment rates, 4–10, 36, 41, 42, 45, 55, 67, 88, 89, 111–120, 122, 196, 260, 645, 1638
- (Uneven) regional development, 611
- Unidirectional, 598, 1512, 1532
- Uniform policy, 1004
- Unit root system, 612
- Universities, 333, 339, 479–482, 486
- Unobserved consumer surplus, 1007
- Unobserved/latent dependent variable, 1621
- Unperceived benefits, 1000
- Unskilled, 115–116
- Untraded interdependencies, 461, 462, 492–493
- Unusual preferences, 25
- U-O. *See* User optimization (U-O)
- Urban, 1050–1055, 1057, 1063
 activities, 763
 change processes, 750, 751
 connectivity, 1080–1081
 economics, 745, 750, 755
 externalities, 484, 485
 hierarchy, 827
 labor markets, 548, 555
 models, 710, 717, 747, 750–756
 planning, 1072, 1073, 1075, 1077–1079, 1081
 policy, 668, 680
 region, 856
 rent gradients, 519
 sprawl, 77, 83–85, 96
 systems, 484, 834, 836–839, 841, 849
 travel choice, 760, 768, 779
- Urbanization, 284, 463–464, 476–482, 484–487, 632, 634, 636, 638, 1095, 1100
- Urbanization economies, 403–404, 409
- Urban sprawl, 77, 83–85, 96, 753, 1044
- USA, 661
- Use benefits, 1006
- User equilibrium (UE), 760, 766, 769, 770, 774–778, 781
- User-generated content (UGC), 731
- User optimization (U-O), 790, 792–799, 802, 808
- Use values, 30, 32
- U.S. regional and state per capita income, 6
- Utility
 curve, 37
 differential, 549–550, 553–555, 559
 function, 66, 70
 levels over space, 20
- V**
- Vaccination, 1323
- Valuation biases, 1004
- Valuation technique, 1031, 1034, 1037–1038, 1043
- Value function, 346–348
- “Value of Statistical Life” (VSL), 995–997, 1007
- Valuing environmental quality improvements, 994

- Variance, 1436–1438, 1440–1453, 1455, 1457, 1458
 Variance-covariance structure, 1347, 1351
 Variance-decomposition, 1444–1446, 1454, 1455
 Variance inflation factors (VIFs), 1444, 1451, 1456
 Variational inequalities (VI), 238, 242–245, 250, 254, 257, 761, 764, 769, 784, 789, 793, 797, 801–808
 Variations in utility over space, 23
 Variety, 458–460, 463, 468, 476–477, 479, 483–487, 540, 541, 544–552, 554–560, 562, 566, 567
 Variety-or ‘composition’ effect, 617
 Variograms, 1207–1211, 1213, 1281, 1289–1291, 1388–1394, 1397, 1462–1471, 1473, 1474
 Vector autoregression (VAR) models, 155, 163
 Vector-borne diseases, 1318
 Verdoorn’s law, 265–266, 268–269, 274–275
 Vertex, 1256–1260, 1262, 1263, 1265, 1268–1270
V
 Vertical dimension, 595
 integration of production process, 916
 intra-industry trade, 916
 product differentiation, 916
 specialization, 913–918
 VGI. *See* Volunteered geographic information (VGI)
 VI. *See* Variational inequalities (VI)
 Vicious cycles, 603–604, 1096, 1097
 VIFs. *See* Variance inflation factors (VIFs)
 ‘Virgin landscape’ assumption, 625
 ‘Virgin market’ idea, 625
 Virtual communities, 735
 Virtuous cycle, 604
 Visual analytics, 1175, 1189–1191
 Volunteered geographic information (VGI), 1119, 1120
 Voronoi tessellation, 1259, 1268
 VSL. *See* “Value of Statistical Life” (VSL)
 Vulnerability, 812, 828
 Vulnerability of low-skilled, 115
 Vulnerable groups, 103, 106
 Vulnerable infrastructure, 1055
 Vulnerable population, 93–106

W
 Wage differentials, 23, 24, 36, 855, 865, 869, 994, 1000, 1004, 1006
 Wage-led growth, 270
 Wages, 18–27, 29–32, 333–334, 336–341, 350, 632–638
 compensation, 997–998
 convergence, 18–20, 23, 24, 32
 differentials, 36
 divergence, 23, 27
 and incomes, 5, 10–12, 14
 Wald (W) or likelihood ratio (LR), 1524, 1529
 Walk, 760, 763, 764, 773–775
 Wall map, 1188, 1189, 1191
 Walras–Leontief production function, 897
 Wardrop, J.G., 790, 792, 795
 Wardrop principles, 791, 792
 Warrick, A.W., 1391
 ‘Wasteful’ commuting, 76, 78–83, 87
 Waste site, 976, 1036, 1043
 Water stress, 1060
 Weak ties, 463
 Web 2.0, 731
 Weighted, 1256–1258, 1260, 1261, 1263, 1265, 1269
 Weighted least squares regression, 1441
 Weighting the values, 29
 Weights matrix, 1285
 Welfare, 478, 855–857, 859, 860, 871, 872
 Welfare benefits of free trade, 912
 Wellbeing, 277–288
 White spaces, 463, 464, 471
 Wikification of GIS, 1120
 Willingness to pay, 138, 975, 976, 990
 Within variance, 394
 Women, 117
 Work effort, 29
 Worker productivity, 632–638
 World city, 1254

Y
 100-Year flood event, 1057
 Young people, 117

Z
 Zero flow magnitudes, 1661, 1662, 1671
 Zipf’s law, 199–200, 815, 817, 827
 Zonation effect, 1157–1164, 1167, 1474
 Zone design, 1158, 1162, 1167–1169
 Zoning, 337–338
 laws, 30
 regulations, 98, 103