

# HANDBOOK OF TRANSPORT GEOGRAPHY AND SPATIAL SYSTEMS

*Edited by*

**DAVID A. HENSHER**

*Institute of Transport Studies,  
University of Sydney*

**KENNETH J. BUTTON**

*The School of Public Policy,  
George Mason University*

**KINGSLEY E. HAYNES**

*Institute of Public Policy,  
George Mason University*

**PETER R. STOPHER**

*The School of Transport Studies,  
University of Sydney*



United Kingdom – North America – Japan –  
India – Malaysia – China

Emerald Group Publishing Limited  
Howard House, Wagon Lane, Bingley BD16 1WA, UK

Third edition 2008. Previous editions 1982, 1988

Copyright © 2008 Emerald Group Publishing Limited

**Reprints and permission service**

Contact: books@emeraldinsight.com

No part of this book may be reproduced, stored in a retrieval system, transmitted in any form or by any means electronic, mechanical, photocopying, recording or otherwise without either the prior written permission of the publisher or a licence permitting restricted copying issued in the UK by The Copyright Licensing Agency and in the USA by The Copyright Clearance Center. No responsibility is accepted for the accuracy of information contained in the text, illustrations or advertisements. The opinions expressed in these chapters are not necessarily those of the Editor or the publisher.

**British Library Cataloguing in Publication Data**

A catalogue record for this book is available from the British Library

ISBN: 978-0-0804-4108-5

Printed and bound in Great Britain by  
CPI Antony Rowe, Chippenham and Eastbourne



Certificate number ..... 1985 .....

Awarded in recognition of  
Emerald's production  
department's adherence to  
quality systems and processes  
when preparing scholarly  
journals for print



INVESTOR IN PEOPLE

## INTRODUCTION TO THE SERIES

Transportation and logistics research has now reached maturity, with a solid foundation of established methodology for professionals to turn to and for future researchers and practitioners to build on. Elsevier is marking this stage in the life of the subject by launching a landmark series of reference works: *Elsevier's Handbooks in Transport*. Comprising specially commissioned chapters from the leading experts of their topics, each title in the series will encapsulate the essential knowledge of a major area within transportation and logistics. To practitioners, researchers and students alike, these books will be authoritative, accessible and invaluable.

David A. Hensher  
Kenneth J. Button

## CONTENTS

### Introduction to the series

v

#### *Chapter 1*

##### **Introduction**

KINGSLEY E. HAYNES, PETER R. STOPHER, KENNETH J. BUTTON and  
DAVID A. HENSHER

1. Introduction	1
2. Recent trends in analysis	2
3. GPS, GIS and other acronyms	4
4. Land use and transportation institutions	5
5. The Handbook	7

### *Part 1. Transport and Geography*

11

#### *Chapter 2*

##### **Recent Developments in US Transport Geography**

WILLIAM R. BLACK

1. Some definitions	13
2. Historical background	13
3. Transport geography today	16
4. Recent developments in the field	18
4.1. Deregulation	18
4.2. Activity analysis	19
4.3. Sustainable transport	19
4.4. Environmental justice	20
4.5. Economic development	21
4.6. Geographic information systems	22
4.7. Network design	23
5. Some concluding thoughts	24
6. Closure	24
References	25

#### *Chapter 3*

##### **Institutions, Land Use and Transportation**

ROGER R. STOUGH

1. Introduction	27
2. Theory and recent changes in land use and transportation	28

2.1. Altered demand for urban transportation and land use	29
2.2. Metropolitan decentralization	30
2.3. Edge cities	31
<b>3. Institutions</b>	<b>32</b>
<b>4. Institutional analysis decentralization in metropolitan regions</b>	<b>35</b>
4.1. First-level institutions: US values and culture and metropolitan land use patterns	35
4.2. Second-level institutions: formal institutions	36
4.3. Third-level institutions: governance institutions	38
4.4. Fourth-level institutions: resource allocation and short-term outcomes	39
4.5. Institutional analysis: conclusions	40
<b>5. Conclusions</b>	<b>41</b>
<b>References</b>	<b>41</b>

*Chapter 4*

<b>Transportation Location and Environmental Justice: A US Perspective</b>	
KINGSLEY E. HAYNES	43

<b>1. Introduction</b>	<b>43</b>
<b>2. TRI geography and environmental justice</b>	<b>45</b>
2.1. Environmental justice and TRI	45
2.2. Geography of environmental justice	47
<b>3. Transportation and environmental justice</b>	<b>49</b>
<b>4. Empirical analysis</b>	<b>51</b>
<b>5. Conclusion</b>	<b>59</b>
<b>Acknowledgments</b>	<b>60</b>
<b>References</b>	<b>61</b>

<b>Part 2. Transport and Spatial Form</b>	<b>65</b>
---	-----------

*Chapter 5*

<b>Transport in the Urban Core</b>	
EVELYN BLUMENBERG and RANDALL CRANE	67

<b>1. Introduction</b>	<b>67</b>
<b>2. Density</b>	<b>68</b>
<b>3. Poverty</b>	<b>70</b>
<b>4. Decentralization</b>	<b>71</b>
4.1. Spatial mismatch	72
4.2. The journey to work	73
<b>5. Summary</b>	<b>75</b>
<b>References</b>	<b>75</b>

*Chapter 6***Economic Development and Transport Hubs**

KENNETH J. BUTTON

1. Introduction	77
2. Airports as hubs	78
3. Regional impacts of hub airport	83
3.1. Primary effects	83
3.2. Income multiplier effects	83
3.3. Tertiary effects	84
3.4. Perpetuity effects	84
4. Empirical analysis	85
4.1. Surveys and questionnaires	85
4.2. Multiplier analysis	87
4.3. Econometric models	88
5. International airline hubs	89
6. Conclusions	94
References	95

*Chapter 7***Transport and Spatial Clustering**

JEAN H.P. PAELINCK

1. Introduction	97
2. Preliminary concepts	98
3. Market approach	99
4. Non-market approach	102
5. Synthesis	104
6. Conclusions	106
References	109

*Chapter 8***Connecting Mass Transit and Employment**

THOMAS W. SANCHEZ

1. Introduction	111
2. Elements of travel demand	112
2.1. Trip purpose	112
2.2. Trip timing	113
2.3. Trip origins and destinations	114
2.4. Trip mode	114
2.5. Available routes	115
2.6. Trip frequency	116
3. Work trip factors	116

3.1. Distribution of trip times and day of week	117
3.2. Direction of trip flows	117
3.3. Modal availability	119
<b>4. Other factors</b>	<b>120</b>
4.1. Network characteristics	120
4.2. Network extensiveness	120
4.3. Network connectivity	121
4.4. Physical access – walking distances	121
4.5. Vehicle ownership levels	122
<b>5. Summary</b>	<b>122</b>
<b>References</b>	<b>123</b>
 <i>Part 3. Land Use and Transportation Modeling</i>	 125

*Chapter 9***Overview of Land use Transport Models**

MICHAEL WEGENER

<b>1. Introduction</b>	<b>127</b>
<b>2. Existing urban land use transport models</b>	<b>128</b>
2.1. Urban change processes	128
2.2. Twenty urban models	131
<b>3. Future urban land use transport models</b>	<b>138</b>
<b>4. Conclusions</b>	<b>142</b>
<b>Acknowledgment</b>	<b>143</b>
<b>References</b>	<b>143</b>

*Chapter 10***Integrated Land Use/Transport Model Requirements**

ERIC J. MILLER

<b>1. Introduction</b>	<b>147</b>
<b>2. A framework for integrated modeling</b>	<b>147</b>
<b>3. Design issues</b>	<b>151</b>
3.1. Physical system representation	151
3.2. Representation of active agents	153
3.3. Representation of processes	154
3.4. Generic design issues	156
3.5. Implementation issues	157
<b>4. Evaluation criteria</b>	<b>160</b>
4.1. Credibility criteria	160
4.2. Feasibility criteria	162
4.3. Usability criteria	162

5. Summary	164
Acknowledgments	164
References	165
<i>Chapter 11</i>	
Lowry-type Land Use Models	167
ALAN J. HOROWITZ	
1. Introduction	167
2. Land use model concept: urban form and land rents	168
2.1. Urban form and land rents	168
2.2. Agglomeration	169
3. Residential location models	169
3.1. Basic form	169
3.2. Population segmentation	171
3.3. Measures of attractiveness	172
3.4. Land constraints	172
3.5. The exogenous workplace	173
3.6. Multimodal applications	173
4. Overview of the Lowry model	173
4.1. Typical data requirements	175
4.2. Anticipated results and cautions	175
4.3. Calibration issues	176
4.4. Equilibrium conditions	176
4.5. Deterrence function parameters	176
4.6. Disutility and the value of time	177
4.7. Definition of basic employment	177
5. Derivation of the Lowry–Garin model	177
5.1. Adjustments to residential attractiveness	179
5.2. Adjustments to service attractiveness	180
6. Iterating a land use model with a travel-forecasting model	180
7. Critique	181
8. Closure	182
References	182
<i>Chapter 12</i>	
Econometric Models of Land Use and Transportation	185
MARCIAL ECHEÑIQUE	
1. Introduction	185
2. Theoretical foundations	186
3. A general model of trade and location	188
3.1. Functional relationships	188
3.2. Spatial relationships	190

3.3.	Estimation of spatial prices	192
3.4.	Changing functional relationships (variable demand coefficients)	194
3.5.	Modeling the transport systems	195
3.6.	Integrated spatial system model	196
3.7.	Policy modeling	199
<b>4.</b>	<b>Applications</b>	<b>200</b>
<b>5.</b>	<b>Conclusions</b>	<b>201</b>
	<b>References</b>	<b>201</b>

*Chapter 13***Introduction to Urban Simulation: Design and Development of Operational Models**

PAUL WADDELL and GUDMUNDUR F. ULFARSSON

203

<b>1.</b>	<b>The context and objectives for urban simulation</b>	<b>203</b>
<b>2.</b>	<b>The design and implementation of an operational urban simulation system</b>	<b>205</b>
2.1.	Assess the institutional, political, and technical context	207
2.2.	Assess the stakeholders, value conflicts, and public policy objectives	210
2.3.	Develop measurable benchmarks for the objectives	212
2.4.	Inventory the policies to be tested	212
2.5.	Map the policy inputs to outcomes	213
2.6.	Assess the model requirements	217
2.7.	Make preliminary model design choices	219
2.8.	Select the modeling approach	222
2.9.	Prepare the input data	226
2.10.	Develop the model specification	227
2.11.	Estimate the model parameters	230
2.12.	Calibrate the model system	231
2.13.	Develop the software application	231
2.14.	Validate the model system	232
2.15.	Operational use	232
<b>3.</b>	<b>Conclusion</b>	<b>233</b>
	<b>Acknowledgments</b>	<b>233</b>
	<b>References</b>	<b>234</b>

*Chapter 14***Evolutionary Approaches to Transport and Spatial Systems**

AURA REGGIANI

237

<b>1.</b>	<b>Introduction</b>	<b>237</b>
<b>2.</b>	<b>Spatial choice and processes: the role of spatial interaction models</b>	<b>238</b>

2.1. Spatial interaction models: the analytical form	238
2.2. Spatial interaction behavior and choice behavior	239
3. Non-linear dynamic processes: the logistic form	241
4. Networks and complexity	243
5. Network complexity	245
5.1. Simple models for complex networks: niche models	245
5.2. Complex models for complex networks	246
5.3. Detecting complexity from data	247
6. Network resilience	247
7. Emergence and self-organized criticality	249
7.1. The concepts of emergence and self-organization	249
7.2. The concept of SOC	250
8. Conclusions	251
Acknowledgments	252
References	252

*Chapter 15*

Transportation and Urban Compactness HARRY W. RICHARDSON and CHANG-HEE CHRISTINE BAE	255
1. Introduction	255
2. Implications of urban economic theory	256
3. Historical evolution	256
4. Interpreting data and the urban scale	257
5. The dynamics of transportation, land use, and urban compactness	258
6. Transit-oriented developments	260
7. New urbanism	261
8. Neighborhood types	262
9. Intertemporal changes	262
10. Dispersal and travel behavior	263
11. Information technology	264
12. International comparisons	264
13. Conclusions	265
References	266

*Chapter 16*

Computable General Equilibrium Analysis in Transportation Economics JOHANNES BRÖCKER	269
1. Introduction	269
2. A primer in CGE analysis	270
3. Transportation in CGE analysis	277
3.1. Goods transport	278
3.2. Passenger transport	280
3.3. Economic equilibrium and transport network equilibrium	281

4. Extensions	281
4.1. Imperfect markets	281
4.2. Dynamics	284
5. An example: the spatial effects of trans-European road networks	284
6. Conclusions	286
References	287

<i>Part 4. Data</i>	291
---------------------	-----

*Chapter 17*

Spatial Data Issues: A Historical Perspective	
---	--

PETER R. STOPHER	293
------------------	-----

1. Introduction	293
2. Traffic analysis zones	294
3. Traffic networks	300
3.1. Bus networks	302
3.2. Micro-networks	303
4. Interactions between zones and networks	303
4.1. Zone size and networks	304
4.2. The use of a GIS as a network platform	305
4.3. Network detail and zone size	306
5. Conclusions	307
References	308

*Chapter 18*

Linking Spatial and Transportation Data	
---	--

BRUCE D. SPEAR	309
----------------	-----

1. Introduction	309
2. GISs and transportation models – a US historical perspective	309
2.1. Origins of GISs	309
2.2. Origins of transportation models	310
2.3. Development of commercial software	311
2.4. TIGER and GISs	311
2.5. The Census Transportation Planning Package and GIS	312
3. Conceptual differences between GISs and transportation models	313
3.1. GIS spatial objects and relationships	314
3.2. Network objects and relationships	316
3.3. Translating between linear spatial objects and networks	317
4. Other transportation data structures	320
4.1. Routes	320
4.2. Linear referencing	321
4.3. Matrices	323

4.4. Dynamic spatial objects	324
5. Conclusions	325
References	326
<i>Part 5. GIS Applications</i>	327
<i>Chapter 19</i>	
The Role of GIS in Land Use and Transport Planning HOWARD L. SLAVIN	329
1. Introduction	329
2. GIS in land use planning	330
2.1. Data development, presentation, and access	330
2.2. Data access	330
2.3. Urban information systems and urban analysis	334
3. GIS in land use modeling	335
4. GIS in transport planning	337
4.1. An overview of GIS-T functionality	337
5. GIS in travel-demand modeling	344
5.1. GIS-T use in modeling: the linkage-integration debate	344
5.2. GIS-T application to modeling activities and components	346
6. Concluding remarks	355
References	356
<i>Chapter 20</i>	
The Role of GIS in Routing and Logistics JOHN C. SUTTON and JOHAN VISSER	357
1. Introduction: why use GIS in routing and logistics?	357
2. GIS routing and logistics capabilities	358
2.1. Vehicle routing/dispatching	361
2.2. Arc routing	361
2.3. Network flow and distribution analysis.	361
2.4. Location and allocation models	362
3. Logistics issues	364
4. Public policy-making	365
4.1. Freight modeling	367
4.2. Spatial studies	367
5. Real-time routing and logistics	369
5.1. From static to dynamic information	369
5.2. Convergence of GIS and location aware technologies	370
6. Software	370
6.1. GIS limitations	371
7. Conclusion	373
References	374

*Chapter 21***GIS and the Collection of Travel Survey Data**

STEPHEN GREAVES

375

1. Introduction	375
2. Use of GIS in travel surveys	376
3. Geocoding of survey data	377
3.1. Automated address matching and GIS	378
3.2. How the automated geocoding process works	378
3.3. Partial matches	380
3.4. Checking of geocodes	381
4. Developing the databases	382
4.1. The reference databases	382
4.2. Developing the target database	383
4.3. Spatial bias and spatial stratification	388
5. Summary and future directions	389
References	390

*Chapter 22***GIS and Network Analysis**

MANFRED M. FISCHER

391

1. Introduction	391
2. Network representation and GIS-T network data models	392
2.1. Terminology	392
2.2. The network data model	392
2.3. Non-planar networks and the turn-table	395
2.4. Linear referencing systems and dynamic segmentation	396
2.5. Lanes and navigable data models	398
3. Vehicle routing within a network: problems and algorithms	400
3.1. The traveling-salesman problem	400
3.2. The vehicle-routing problem	402
3.3. Constrained shortest-path problems	405
4. Concluding remarks	407
References	407

*Part 6. GPS Applications*

409

*Chapter 23***Defining GPS and its Capabilities**

JEAN WOLF

411

1. Introduction	411
2. The Global Positioning System	412

2.1. Overview of GPS	412
2.2. PVT determination	414
2.3. Other GNSS	416
2.4. GPS user technologies	417
2.5. GPS receiver output	417
2.6. GPS performance measures	419
2.7. Standalone GPS position accuracy and augmentations	423
2.8. Free satellite-based augmentation systems	425
2.9. GPS modernization (or GPS III)	426
<b>3. GPS capabilities for transport</b>	<b>427</b>
3.1. Highway, transit, airport, and seaport traffic control and security	427
3.2. E911	427
3.3. Location-based services	428
3.4. Combined measures of travel, physical activity, and health	428
3.5. Mobile source emissions analysis and modeling	430
3.6. Long-term travel studies	430
<b>Appendix: Internet resources for GPS</b>	<b>430</b>
<b>References</b>	<b>431</b>

*Chapter 24***GPS, Location, and Household Travel****PETER R. STOPHER**

<b>1. Introduction</b>	<b>433</b>
<b>2. GPS as a solution</b>	<b>434</b>
2.1. Types of GPS device	435
2.2. What GPS can do	438
2.3. What GPS cannot do	440
<b>3. Processing GPS data</b>	<b>441</b>
3.1. Problems with GPS data	442
3.2. Accuracy of GPS	443
3.3. Wearable GPS devices	444
<b>4. The future of GPS</b>	<b>445</b>
4.1. Privacy	447
4.2. Respondent burden	447
<b>5. Conclusions</b>	<b>448</b>
<b>References</b>	<b>449</b>

*Chapter 25***GPS and Vehicular Travel****GEOFF ROSE**

<b>1. Introduction</b>	<b>451</b>
------------------------	------------

2. Key technology links and applications	452
3. Remote monitoring of vehicle location	455
4. Arrival time information	456
5. In-vehicle navigation	457
6. Intelligent speed adaptation	459
7. Advanced driver assistance systems	460
8. Electronic payment and charging	461
9. Unresolved issues	463
9.1. Map database related	463
9.2. Human factor considerations	464
9.3. Willingness to pay	464
9.4. Managing privacy	465
9.5. Public and user acceptance	465
10. Conclusions	466
References	466

*Chapter 26*

Traffic Monitoring Using GPS CESAR QUIROGA	469
1. Introduction	469
2. Measuring travel times, speeds, and delays using GPS	471
2.1. Generating routes, checkpoints, and segments	471
2.2. Linearly referencing GPS data	473
2.3. Calculating segment travel times, speeds, and delays	474
2.4. Calculating intersection delays	477
3. Data management	479
3.1. Architecture	479
3.2. Linear referencing and computation of travel time	482
3.3. Intersection delay	485
4. Summary	487
References	487

*Chapter 27*

Other Transportation Applications of GPS SHAUNA L. HALLMARK	489
1. Introduction	489
2. Centerline mapping	490
3. Inventory management	492
3.1. General	492
3.2. Mobile mapping systems	493
4. Automatic vehicle location	493

4.1. In-vehicle navigation systems	494
4.2. Fleet management	494
4.3. Concept winter vehicle	495
5. Safety	495
5.1. Crash location	495
5.2. On-board crash notification systems	496
6. Locating environmentally sensitive features	497
7. Summary	497
References	498
 <i>Part 7. Spatial Cognition</i>	499
 <i>Chapter 28</i>	
Cognitive Maps and Urban Travel	
REGINALD G. GOLLEDGE and TOMMY GÄRLING	501
1. Introduction	501
2. Basic concepts	501
2.1. Cognitive maps	501
2.2. Cognitive mapping	502
3. Transportation issues	502
3.1. Cognizing transportation networks	502
3.2. Travel behavior	504
3.3. Path selection criteria	506
3.4. Navigation and wayfinding	506
3.5. Route learning	507
3.6. The role of trip purpose	508
3.7. Travel guidance	508
4. Incorporating cognitive maps into travel choice models	509
5. Conclusion	510
References	511
 <i>Chapter 29</i>	
Spatial Processes	
RYUICHI KITAMURA	513
1. Introduction	513
2. Trip-based studies and their limitations	515
3. Trip-chaining analyses	518
4. Classification approaches	522
5. Simulation approaches	524
References	528

*Chapter 30***Mental Maps**

LISA WESTON and SUSAN HANDY

533

1. Introduction 533
2. What are mental maps? 535
3. How do people create mental maps? 536
4. How have mental maps been used? 540
5. How can transportation professionals use mental maps? 543
6. Conclusions 544
- References 544

*Part 8. Geosimulation*

547

*Chapter 31***Geosimulation, Automata, and Traffic Modeling**

PAUL M. TORRENS

549

1. Introduction 549
2. Recent developments in the research landscape 549
3. The emerging geosimulation approach 550
4. Automata as geosimulation tools 552
5. Modeling vehicular traffic 554
  - 5.1. Spatial topology 554
  - 5.2. Entity descriptions 555
  - 5.3. Neighborhood definitions 555
  - 5.4. Time 556
  - 5.5. Rules 556
6. Modeling pedestrian traffic 557
  - 6.1. Entities 558
  - 6.2. Spatial topology 559
  - 6.3. Time 559
  - 6.4. Neighborhoods 559
  - 6.5. Rules 560
7. Conclusion 561
- References 562

*Part 9. Networks*

565

*Chapter 32***Design and Analysis of Transport Networks**

HAI YANG and XIAONING ZHANG

567

1. Introduction 567
2. Formulations of network design problems 569

2.1.	General framework of the bi-level model in network design and analysis	569
2.2.	The continuous network design problem (CNDP)	570
2.3.	The optimal toll pricing problem (OTPP)	571
2.4.	The signal-setting problem	571
3.	Properties of the bi-level model and the solution algorithm	573
3.1.	Non-differentiability of the reaction function	573
3.2.	The marginal-function-based solution algorithm	574
4.	Applications in location choice, land use, and network capacity	580
5.	Conclusions	580
	References	581

*Chapter 33*

<b>Spatial Equilibration in Transport Networks</b>		
ANNA NAGURNEY	583	
1.	Introduction	583
2.	Basic decision-making concepts and models	585
2.1.	System optimization versus user optimization	586
3.	Models with asymmetric link costs	591
4.	Multiclass, multicriteria traffic network equilibrium models	599
4.1.	Traffic network equilibrium conditions	603
5.	Dynamics	604
6.	Summary and new directions	606
	Acknowledgments	606
	References	606

*Chapter 34*

<b>Traffic Assignment Methods</b>		
WILLIAM H.K. LAM and HONG K. LO	609	
1.	Introduction	609
2.	Route choice principles	610
3.	Three traffic assignment models	612
3.1.	Deterministic UE model	612
3.2.	Logit-based SUE model	613
3.3.	Probit-based SUE model	615
4.	Case study	618
4.1.	Model calibration	620
4.2.	Model validation	621
5.	Concluding comments	623
	Acknowledgment	623
	References	624

<b>Part 10. Time Use</b>	627
<i>Chapter 35</i>	
Time Use and Activity Systems	
ANDREW S. HARVEY	629
1. Introduction	629
2. Activity systems approach	629
3. Time use and travel behavior	631
3.1. The time use perspective	631
3.2. Travel perspective	632
4. Time use measurement	633
4.1. Time use data collection methodology and instruments	637
4.2. Activities and context	638
5. Time use analysis	640
5.1. Unit of analysis	640
5.2. Activity measures	642
6. Advantages and challenges of the time use approach	644
7. Conclusions	645
References	645
<i>Chapter 36</i>	
Activities in Space and Time	
HARVEY J. MILLER	647
1. Introduction	647
2. Time geography	648
2.1. Activities in space and time	648
2.2. Space-time path and prism	649
2.3. The individual and space-time aggregates	651
2.4. Contrasts with time use and activity analysis	651
2.5. Time geography and transportation research	652
3. Information technologies and the new time geography	652
3.1. Representation of space-time environments	653
3.2. New methods for data collection	655
3.3. New methods for data analysis	656
3.4. Extending time geography to cyberspace	657
4. Conclusion	658
References	658
Author Index	661
Subject Index	689

*Chapter 1*

## INTRODUCTION

KINGSLEY E. HAYNES

*George Mason University, Fairfax, VA*

PETER R. STOPHER

*University of Sydney*

KENNETH J. BUTTON

*George Mason University, Fairfax, VA*

DAVID A. HENSHER

*University of Sydney*

### 1. Introduction

Historically it is impossible to separate human geography from transportation. Even in prehistoric times the ability to populate large parts of the globe was only possible because of the existence of long defunct land bridges. Modern civilization grew at the cross-roads of major trade routes, often involving seaports, the hubs of the maritime highways of the day. More recently, urbanization and sub-urbanization have been inextricably tied to developments in mechanized transportation, first local railways and trams, and then mechanized road transport. Globalization has been growing, and there is ample evidence that technologically advanced shipping and air transport and the development of unitization have been as much – if not more – a driving force as reforms of international treaties. For anyone who has visited the ports of Singapore and Hong Kong, for example, it is difficult to envisage how the volume of trade they handle would be possible without the invention of the container, and without this trade the cities of South-East Asia would not have the form we know today.

The intellectual interest in the links between transportation and land use has an impressive history over the past 150 years. The foundation of much of modern microeconomics can be traced back to the efforts of von Thünen in the mid-nineteenth century to explain agricultural land use patterns around free-standing towns and cities. A framework later developed by William Alonso in the context

of rent levels of concentric rings of activities around concentric city centers. Transportation systems were also important in the pioneering work on city systems and, in particular that of Christaller in the 1930s, and later Lösch in the 1950s on central-place theory. The use of the gravity model framework has become widespread in the modeling of migration and the concentration of populations, with the quality of transportation or travel time being used as the impedance factor.

However, while there has been a significant involvement in transportation by geographers and regional scientists for at least the past century and a half, the geographic aspects of transportation only became much more of a central focus at the end of the twentieth century. The development of new urban economics in the 1970s, with its emphasis on rigorous mathematical modeling, and more recently new economic geography with its heightened focus on trade and scale effects, have brought to the fore the importance of transportation in understanding spatial land use patterns. At the more pragmatic level, there has been enhanced interest in the role transportation plays in facilitating trade and national development in the context of a globalized economy, and in the role that it may play in the consolidating macro spatial political-economic entities such as the EU and the North-American Free Trade Area (NAFTA).

## 2. Recent trends in analysis

It is now accepted that transportation is a geographic phenomenon; indeed, according to *Webster's Dictionary*, geography is defined as the study of 'the earth's surface: a science that deals with the earth and its life; especially the description of land, sea, air, and the distribution of plant and animal life including man and his industries'. Furthermore, human geography is especially concerned with the distribution of human settlements, the ways in which such settlements develop, and all of the implications contained within that. It is also very much concerned with spatial systems, and all things related to such systems. Transportation is a significant spatial system, and one that has substantial influences on the ways in which human settlements evolve.

The study and analysis of transportation phenomena is essentially a map-intensive process. However, through much of the computer era of transport analysis, computer representations of maps have been largely restricted to drawing straight lines between spatially located points, to represent transportation networks, and the production of manual maps on to which have been superimposed various pieces of information relating to such things as population distributions, the distribution of certain demographics of the population, and the products of the transport-planning models themselves.

Over perhaps the past 30 years, there have been technological changes in the way transportation and geography have been studied that have come together to produce a transport geography that builds on past issues, such as location, land use, infrastructure, and modeling, but using data and information in ways that were not possible previously. The difference is that not only can real-time geocoded knowledge of transport–spatial systems be integrated into analysis but also it can be used to intervene directly in transport flows on a network through direct vehicle guidance. The technology to do this includes the Global Positioning System (GPS), used for tracking movement on to a geographic information system (GIS), which can integrate and respond to rapidly changing and erratic movements such as network flows and congestion build up, and slower but continuously changing processes such as land use and infrastructure adjustments. Further, these observations can be communicated directly and precisely to transport participants – passengers, drivers (individual and mass transit), and potential participants in real time through intelligent transportation systems (ITS) or telematics.

These new and revolutionary technologies focus on the spatial organization of transportation movements with the opportunity for real-time guidance, advising, and intervention in transportation processes. However, even with these advances, fundamental interdependencies of transport and society remain centrally important to land use and traffic generation; urban design and travel demand; emissions, congestion, and air quality; local and community decisions relative to regional and national priorities; and mobility considerations versus environment concerns.

In the chapters that follow, the contributions to this handbook clarify the impact of these technological changes on spatial systems, with the intention that transportation analysts will become more comfortable with the new geographical information system opportunities. The aim also is that they will appreciate the continuance of fundamental geographical constraints. The contributions are designed to clarify for geographers and planners the advanced use of spatial systems technologies in the field of transport in order for them to appreciate the new questions that need to be addressed. For example, now that intervention is possible in real time, what are the likely socio-behavioral responses to sender information and intervention? What intervention strategies are possible and when should they occur? How will drivers react to information overload? How will these technologies affect the form and density structure of the city? Will ITS/telematics affect the amount and nature of travel, and what will the impact of this be on the environment?

The world of transportation has changed. New opportunities have been created. The fundamental questions continue to be important but now a series of new questions need to be addressed as well. We focus, in this handbook, on spatial system technologies, GPS, GIS, mobility management, real-time data system

integration, and their consequences and opportunities for transport system mobility and management. We do not provide much background on the integration of the different elements of ITS/telematics that have accelerated the incorporation of these spatial system technologies into the geography of transportation. This material has been covered in Volume 3 in this series (*Handbook of Transport Systems and Traffic Control*).

### 3. GPS, GIS and other acronyms

Traditionally, empirical analyses at the interface between geography and transportation had to rely on infrequent census data and occasional surveys. Recent years have seen major developments in the information available for studying the key linkages between transportation and land use. The Global Positioning System is the name given to a system developed and deployed by the US Department of Defence in the late 1980s, and made available for public use during the early 1990s. The system was initially deployed for use by the armed services for positioning in relation to military tactics and needs. When it was made possible for the general public to make use of the system this entailed 'selective availability' – a disturbance in the accuracy of the signal, making the best accuracy of an uncorrected GPS signal around  $\pm 100$  m. Selective availability was turned off in 1999, and it then became possible for civilian applications of GPS to achieve accuracy, without use of secondary correction procedures, of  $\pm 10$  m or so. This offers much more sophisticated information, not only for the immediate user but also for subsequent study and analysis.

Geographic information systems are systems of relational databases that are geographically referenced and stored, and permit a wide range of mapping and charting to be performed. A GIS is made up, essentially, of map layers, each of which may contain points, lines, or polygons. These represent three different ways in which geographic information can be represented. Points represent features that exist at a single location, such as a mountain summit; a building, monument, or similar man-made object; or the location of a GPS receiver at an instant in time. Lines represent features that extend over distance, but that have relatively nominal width, such as roads, rail lines, rivers and streams, and footpaths. Polygons represent areas, such as lakes, oceans, river estuaries, parks, and shopping centers. Each of these representations of the location of a feature on the earth's surface can have associated with them certain data. These data are stored in accompanying databases, which have a reference to the mapped feature, such as latitude, longitude, and altitude. There are few limitation as to how much information can be stored about any specific feature – limits are generally imposed only by the size of the computer and the supporting database software.

The principal advantages of a GIS are that different layers of information can be superimposed over each other to create composite maps, and a wide variety of pieces of information can be associated with the layers of these maps. Each layer can be visualized as a transparent sheet, except for the points, lines, or polygons that are part of that layer. These features may be colored, or may exist only as outlines that allow detail in lower layers to remain visible. By first laying down a map showing, for example, the coastline, and then superimposing over this a line layer containing rivers, a line layer containing roads, another containing rail lines, etc., and then possibly adding a layer showing parks and forests, and a point layer showing landmark features, the result is a map of a region that most geographically inclined people would have no trouble in reading. Labels describing the data for any given layer can be added to the map. In addition, various charts and other options can be displayed on the map, or built up as a series of pictures surrounding the map.

The capabilities of GIS are enormous in conveying a wealth of information in a visual and pictorial manner, enormously simplifying the communication of that information. Although the initial development of GIS was to support environmental and land information applications, the application of these systems to transportation was an inevitable step, and this took place mainly during the 1980s and early 1990s.

Another technological development of interest, although yet to be fully exploited in the transportation analysis field are Global System for Mobile Communications (GSM) mobile telephone systems. These can be used, for example, to plot the locations of transmissions by measuring the time it takes for a signal to travel between a base station and the mobile telephone and back. An added value of GSM may be in hybrid systems that combine GSM technology with other communication and positioning systems. A combined GSM-GPS system, for example, would have benefits over GPS technology used alone: GPS systems do not work well in areas where there are many tall buildings, or inside some vehicles, and do not work at all in tunnels; GSM systems, on the other hand, are much less affected by tall buildings, work well inside virtually any vehicle, and usually work in tunnels. Therefore, hybrid systems offer the potential of a more complete positioning capability, even though current GSM technology does not provide a level of accuracy that is comparable to GPS.

#### 4. Land use and transportation institutions

There is little debate that land use and transportation interact with each other. Transport is necessary to open up land for development. In turn, unless there is use made of land, there is no reason to provide transportation connections to it. Interestingly, however, land use and transport have, for the past century at least,

been disconnected in almost every sense with respect to management and planning. Planning of land use is predominantly located in agencies that are involved in development planning, while transportation is the purview of agencies that are focused on the transport task.

Further, decisions on land use are often made in relation to concerns with development densities, production of development-based revenues, and notions of densification of development, and infill of vacant land. Very often, these decisions are made without much consideration of the ability of the existing transportation infrastructure to handle the demands that the development will create. Similarly, decisions to release vacant land on the outskirts of many urban areas are also often done without any real understanding of, or planning for, the transportation implications of those developments.

At the same time, those who are responsible for planning the transportation infrastructure are frequently unaware of the direction in which land development is moving, or, often for political reasons, are not allowed to take into account probable future development that may place additional demands on the transport infrastructure. Transportation planners often find that they have to be reactive to the decisions made by land use planners, and, with increasing frequency, find themselves in an impossible situation where options for increasing transport capacity to meet demands generated by new development simply do not exist at any reasonable price. In addition, the funds for expenditure on transport are often not matched to the needs being generated from growth in residential and commercial land uses. This situation is often made worse when the physical jurisdictions of the land use authorities do not coincide with those of the transportation planning agencies, or when the funding structure for transportation actions does not fit either the temporal or physical parameters of land use developments.

One of the reasons for this traditional disconnection between land use and transportation has been a lack of a sound common geographic base to use for integrated planning of land use and transport. Certainly, the development of capabilities such as GIS should help eliminate this cause of a disconnection between land use and transportation. On the other hand, part of the disconnection probably stems from a combination of political issues and organizational structures that continue to make it difficult to integrate land use and transport within a political and organizational framework. Institutional inertia is one of the stronger forces that plays a role in the disconnection.

With the increasing incorporation of geography into transport planning, efforts at integrating land use and transport planning, and even environmental assessment, are receiving a new impetus, which is reflected in the chapters of this handbook, although this process is slow. Despite the fact that GIS and GPS systems have both been around for nearly 20 years, there are still many urban areas throughout the world that have no formal land use modeling capability, and

for whom integration of land use and transport planning is not even a pipe dream. The urban areas where an integrated land use and transport model is in place and is used are still very much in the minority.

## 5. The Handbook

This is the fifth volume in the *Handbooks in Transport* series. The earlier volumes have been concerned with transport modeling; transport logistics and supply chain management; traffic systems and traffic control; and environmental aspects of transport. This volume concentrates on the geographic aspects of transportation. As with others in the series, this volume is not intended as a textbook or research monograph, nor does it attempt to cover the topic area of transport and geography exhaustively, or in great depth. It does attempt, though, to provide a useful manual or guidebook to the topic area, and to offer a wide range of views on transport and geography that may lead the reader to enquire more deeply into specific areas that are touched on herein. It is intended that this handbook should be as accessible as possible, especially to those who are relatively unfamiliar with this area of work.

Some areas are not covered in this handbook because they are dealt with in other volumes in the series, such as the area of ITS in the *Handbook on Transport Systems and Traffic Control* (Volume 3) and that of the environment in the *Handbook on Transport and the Environment* (Volume 4). However, the material in this volume has been chosen to provide an overview of the area of spatial systems in transportation, and the chapters have been grouped so as to reflect the major elements of this broad sector. As with other handbooks in this series, the chapters are all original, and the international collection of authors were again selected for their knowledge of a subject area, and their ability to communicate basic information in their subject area succinctly and accessibly.

The chapters in Part 1 of this volume provide a broad overview of the spatial aspects of transport, review the history of transportation geography, and the continued role of theory in transport, and the related evolution of urban centers.

Part 2 outlines the links between transport and spatial form, and explore particularly the ways in which transport has and does influence urban form, and the principal ways in which transport is an enabler and a shaper of economic and spatial growth.

The chapters in Part 3 in many ways represent a disaggregation of the preceding chapters. These chapters have a strong orientation toward looking at the different paradigms that have been used to model land use, and the interactions of land use changes and transportation. They form a backdrop to Part 4, which considers spatial data. One of the most important common features of transportation/land use models is that they require a large amount of data. Land use models are perhaps unique in the spatial transportation sphere in that they really cannot

operate easily on sample data – simply because it is not possible to create such a thing as a representative sample of land use patterns. As a result, most land use models require comprehensive land use data on the entire region to which the models are applied. This results in a significant concern with the data aspects of spatial systems and transportation. Integration of spatial data with sample transport data also becomes an issue.

With the focus now on data, Part 5 looks at GIS and applications of GIS in various areas of transport analysis and modeling. In these chapters, the authors explore the role and use of GIS in land use and transport planning, routing and logistics, travel surveys, and network analysis. This treatment then leads into chapters that look at the opportunities afforded by GPS systems to enhance the spatial information used in transport analysis and modeling. These chapters explore the potentials of using GPS in a variety of applications relating to the spatial aspects of transport. The three areas of data, GIS and GPS applications are of central importance for any appreciation of what has happened technologically to the field of geography, and why it is increasingly important to operational concerns of transportation analysts and transport managers. Spatial data, and their display through layer-by-layer integration via GIS referencing and real-time network link-by-link management, are the bridgehead to the new geography and are central to transport analysis. Similarly, GPS systems, utilized in household vehicle traffic and freight flow management, have opened up a new set of opportunities and responsibilities in transport integration.

Part 6 looks at the issue of how people perceive spatial information, space–time processes, and mental maps.

Spatial cognition, leading to the potential behavioral response of individuals to how things are organized in space, and how people respond to the delivery of spatial information, is central to a whole series of questions related to intervention in transport flow patterns, and is discussed in Part 7.

Part 8 investigates the area of geosimulation, which is an outgrowth of the data, GIS, and GPS techniques that provide the spatial information underpinnings for this type of analysis. However, geosimulation is also dependent on the spatial cognition area, because it is there that the appropriate processes to be simulated are identified.

Part 9 deals with transportation networks. This is an important concept in transport geography. It recognizes that channeling is central to part or all of mobility. Channeling takes place over designed networks that are often abstracted as infrastructure. Sorting out the flows over such networks is central to any economy, and is the process of transport equilibration. However, transportation networks are often redundant and robust, providing multiple-path systems that create spatial interdependencies. Of course, one of the most important elements in all network-equilibrating processes is time utilization, which leads into the final chapters in this handbook, comprising Part 10. These chapters look at

relationships between time, space, and activities. The space–time relationship in transport systems can only be fully appreciated when disaggregated geocoded data are used. This has moved this central area of concern forward, as time use, spatial patterns, and network structure can be increasingly integrated into space–time models of transport activity.

In summary, the reader is introduced to the new spatial system technologies that are bringing geography and transport management and analysis together to produce a new transport geography.

***Part 1***

**TRANSPORT AND GEOGRAPHY**

## RECENT DEVELOPMENTS IN US TRANSPORT GEOGRAPHY

WILLIAM R. BLACK

*Indiana University, Bloomington, IN*

### 1. Some definitions

Geography is that field of science concerned with the distribution of phenomena at or near the earth's surface. As a discipline it studies both human and natural features and patterns and seeks to explain why these are located where they are. The phenomena may represent more or less permanent features on the earth's surface, e.g. rivers, highways, or cities, or the phenomena may represent temporally changing attributes, e.g. migration, crop yields, or forest fires, aggregated into some meaningful temporal units of analysis.

Transportation is concerned with the movement of goods and people between different locations and the systems used for this movement. Included in the former would be the journey to work, trade flows between nations, commodity flows within a single nation, passenger flows by various modes, and so forth, and those factors that affect these flows. In general, movement within a single industrial firm or building, or the migration of population, is not included in this area.

Transportation or transport geography may be defined as that discipline concerned with explaining the location of transportation facilities and the magnitude of flows over or near the earth's surface: these two foci may be represented by network analysis and flow analysis. Other definitions are possible, but the focus on networks and flows is common to these as well. For many years, transport geography was viewed as part of economic geography, but ties to this field are not nearly as strong as they once were.

### 2. Historical background

The focus here is primarily on the development of transport geography in the USA, although similar paths were followed in the development of the field in

other nations. The US paradigm of a quantitative analytical approach to the subject was followed by practitioners in the UK, Sweden, Germany, Australia, and Canada. More traditional approaches were also retained in some of these nations. For example, the UK maintains a much stronger modal orientation than is typical of the USA, but it is evident that the quantitative analytical paradigm has a strong hold even there. So while the focus is on US transport geography, much of what is said here is applicable to the discipline in other countries. At the same time, the influence of UK geographers on this paradigm has been very important (e.g. Wilson, 1967, 1974; Haggett and Chorley, 1969).

A review of recent textbooks would suggest that work in this area may have begun in the 1800s, but this would be the wrong impression. It is true that work done in the 1800s may be relevant to looking at some of the problems of interest today, but it was not viewed as transport geography at that time. This was primarily work performed by transportation engineers involved in locating transport routes, specifically railroads, and trying to generalize their location principles to other situations (e.g. Lalanne, 1863; Wellington, 1888).

During the 1920s and 1930s some geographers became involved in describing the patterns of railroads and airline routes, and some of this research may be viewed as the earliest geographical work in this field. For the most part, this research was very descriptive – there was no attempt to generalize findings – and as a result most of it can be set aside. It is discussed today only in the development of the field context as it is being discussed here.

In the post World War II era a number of potential transport research problems began to emerge that were of interest to geographers. Among these were problems related to the growth of the automobile and its use, declining revenues from rail passenger service, the growth of air transport, and traffic congestion in cities. In addition, there was a growing concern with what was moving on the railroads of the USA. This resulted in the collection of a sample of commodity flow data for the major railroads operating at the time: the carload waybill sample.

Simultaneously, some social and physical scientists began to work with analog models for the movement of goods and people. It was observed that the movement of people tended to follow a ‘gravity principle’ wherein the interaction between places was directly proportional to the product of their masses (populations) and inversely proportional to the square of the distance between them. This tendency was operationalized in the gravity model, which did a reasonable job of replicating observed flow patterns, and this led to the development of the short-lived field known as social physics. The notion that the movement of goods and people behaved in much the same manner as inanimate objects contained the seeds of its own destruction, and the field soon disappeared, but the gravity model survived.

Surveys of the field of transportation geography in the 1950s reveal an eclectic range of research being done in the field at that time. One thing that is clear from reviews and textbooks written early in that era is their heavy emphasis on

the economic aspects of transportation. To some extent this is understandable. Although the growth of the automobile was apparent, there was little or no national-level data to reveal exactly how bad the problem of congestion was becoming. Transit operations were beginning to fail in the USA, but there was little data to confirm even this. Air transport was beginning to grow, but jet aircraft would not arrive in commercial service until late in the 1950s. The air passenger service was still small enough not to generate concerns, but there were data available on these flows. The railroad passenger service was in generally poor condition, and the industry, for the most part, wanted to drop it altogether. Freight rates were available, usually published as tariffs, and therefore there was much writing about freight rates and their influence. Data on port activity were also available, and ports became the focus of many studies at this time. Transport geography could hardly be viewed as an organized discipline in the 1950s, and it was usually studied as a subfield of economic geography because most of the data available were of an economic nature.

This changed in the 1960s due to several unrelated events. One of the first of these stemmed from the Federal-Aid Highway Act of 1956. This act funded a major US highway expansion program that would connect the major metropolitan areas of the USA in a single network, formally known as the National System of Interstate and Defense Highways. This program led to the recognition at the federal level of a need for urban transportation studies. In the following decade more than 200 such studies were undertaken. Among the early US studies estimates of future traffic growth were developed based on crude techniques and minimal information. Some planners wanted something a little more rigorous than this, and geographers and regional scientists responded with the gravity model. Data collection efforts for the urban transportation planning process, i.e. data on traffic generation (production and attraction of traffic for zones of a city), trip distribution (connecting trips produced to areas of attraction), modal split (breaking the traffic up and assigning it to available modes), and traffic assignment (assigning traffic to specific routes of the transport network) were all tightly tied into the use of the gravity model, and that model became well known throughout the geography, planning and engineering literature in a short time. There were numerous other flow and distribution models offered at the time that had a more logical foundation, notably the intervening opportunities and the abstract mode models, but the federal government was paying for most of these studies and advocated the use of the gravity model (Bureau of Public Roads, 1965).

A second event was Northwestern University receiving a major research contract from the US Army. This project led to several geographers becoming involved in the application of quantitative methods to transportation-related problems. Some of the earliest applications of graph theory as a way of describing networks and the introduction of connectivity concerns grew from this project.

Explaining the location of transportation routes, i.e. the network generation problem, procedures for grouping flows using factor analytic methods, as well as concerns for optimal networks grew out of this project and had a profound effect on transport geography for more than a decade as it led to a paradigm shift with a focus on network and flow analysis.

A third event was actually generated by the first. Highway construction had led to major impacts on communities, air quality, and transit operations, and transport geographers began working in some of these impact areas as well. Choice of mode and other behavioral considerations were of interest here, and this resulted in a significant "human" addition to the literature that some thought had become too quantitative and in other cases too abstract.

To be sure, these events did not occur in a vacuum. The interstate highway program was a massive financial undertaking, and concerns over impacts on different population groups were ushered in with concerns over civil rights generally. There was a quantitative revolution occurring in the social sciences, and this certainly assisted the researchers at Northwestern University. The result of all this was the development of transport geography as a fundamental research area, with some of the most interesting geographical research of that era being undertaken. The end result of these various events and stimuli was the development of a coherent body of research.

### **3. Transport geography today**

The field of transport geography today may be viewed as consisting of the following topical structure or framework, comprising 18 areas:

- (1) Transport history as it relates to the current spread and distribution of the major transport networks on the earth's surface.
- (2) A discussion of current systems, their networks, and the types of flows that take place on these systems.
- (3) The basic elements of networks used primarily for analysis purposes; this would include an introduction to graph theory.
- (4) Measurement of networks, including notions related to accessibility, connectivity, structure, and so forth.
- (5) Location theory as it relates to the placement of transportation routes and networks.
- (6) Trade and commodity flows, including material ranging from the bases for interaction to new trade theory.
- (7) Methods of flow analysis, including basic flow models, optimal flow systems, factor analytic ideas, and network autocorrelation.
- (8) The prediction problem of flow generation, which has as its focus models of production and attraction for goods and people.

- (9) Spatial interaction and the gravity model, including models ranging from the unconstrained model to the fully constrained model, as well as other flow models.
- (10) Spatial choice models concerned with the choice of destination, mode, and routes.
- (11) Transport policy from a geographic perspective, which attempts to demonstrate why public sector policy positions differ geographically.
- (12) The urban and regional transport planning or forecasting process with particular emphasis on its basic geographical attributes.
- (13) Transport impact analysis, which seeks to illustrate the geographical nature of many of these impacts.
- (14) Transport and its environmental impacts – worthy of particular emphasis – where the environment is viewed as consisting of humans, soils, water, geomorphology, animal life, vegetation, the climate, and the atmosphere.
- (15) Transportation, the economy, and economic development, including models of transport development and transport impacts on the economy; concerns over transport costing and the use of metrics other than distance and travel time to reflect this are also included here.
- (16) Trends, including institutional changes, since many of these have impacts on the transport sector; typical topics would include globalization, deregulation, and privatization.
- (17) Congestion – viewed in many circles as the most significant of our transport problems today; it is the problem least tractable, and since it is essentially a spatial clustering problem on a network, transport geographers need to be more involved in its analysis (Garrison and Ward, 2000).
- (18) Sustainable transport – an area of substantial interest today; it has global significance with a focus on markets, resources, and travel and transport demand.

The above 18 areas should be the major focus of transport geography today from this author's viewpoint. However, there is virtually nothing above on transport modes as such or the study of ports and shipping, and certain aspects of these, e.g. shipping alliances and containerization, are very important. These could probably be included as part of the trends in topic 16 of the framework. Obviously not all of the possible topics have been listed here. However, the framework does identify a general analytical orientation to the field that should apply to any area or mode, and there are very few other topics that would not fit within the structure proposed.

Before proceeding to examine recent developments in these areas, it should be noted that certain of the topical areas noted above do not change very much or very often, e.g. the use of graph theory and some of the basic measurements of

networks have been in use for 50 years and probably will not change much in the future. Therefore, the discussion that follows will focus on those areas that have seen the most development.

#### **4. Recent developments in the field**

Recent developments will be viewed here as events occurring within the last 15 years or so. The general areas examined here include deregulation, activity analysis, sustainability of transport, environmental justice, economic development, geographic information systems (GIS), and network design.

##### *4.1. Deregulation*

The bankruptcy of the Penn Central Transportation Corporation and seven other railroads in the midwest and north-east of the USA in the early 1970s led to the largest corporate reorganization ever undertaken in that country to that point in time. It resulted in the passage of the Regional Rail Reorganization Act of 1973 and the Rail Revitalization and Regulatory Reform Act of 1976, and these resulted in the Consolidated Rail Corporation being created from the more viable parts of the former railroads. In general, the US House of Representatives and the Senate do not particularly favor enacting what could be viewed as essentially regional legislation, and this is probably why the Staggers Rail Act was passed in 1980. This act deregulated the rail industry, giving individual operators considerably more freedom to acquire properties, abandon track, and set rates.

During the above rail deregulation activity the Airline Deregulation Act of 1978 became law. This heavily regulated sector prior to deregulation now found it possible to drop services, add new services, and adjust fares without the need for regulatory agency permission.

Deregulation of US motor carriers and intercity bus carriers came next, with legislation enacted in 1982. Any individual who could demonstrate some level of financial responsibility could form a trucking company.

Transport geography research in the general area of deregulation has focused on the rail and airline sectors. With regard to the former there has been a significant amount of research on rail abandonment and its impacts on communities. The airline sector has seen appreciably more research, with some of the best of this research demonstrating the fallacy of US airline deregulation as it was originally proposed. For example, Goetz (2001) notes that the deregulators assumed there were no economies of scale in the industry, that barriers to entry were insignificant, and that competition would prevent monopoly pricing. His research demonstrated that none of these assumptions held.

Surprisingly, there has been little work in the motor carrier deregulation area in the USA. This is partly due to the proprietary nature of information for this mode. Most companies do not want to reveal the losses they are suffering or the customers they have lost. Nevertheless, data on bankruptcies among middle level trucking firms and the loss of carriers moving LTL (less than total load) traffic have been substantial. In the latter case, only four of the 50 US companies in this sector in 1980 are still in operation at the time of writing. Information is available on the growth of motor carrier owner-operators during this period, and these data can be examined.

#### *4.2. Activity analysis*

The general research area of activity analysis dates from the 1970s, with substantial quantities of research appearing in the early part of the 1980s. The initial idea was that certain trips, notably those to work, were temporally constrained and spatially invariant. A closer look revealed that this was not the case, and numerous researchers quit working in the area. This was perhaps a bit too radical because there are clear indications that travel in total is constrained temporally. There is only so much time that people can spend traveling to work, to shop, and so forth in a typical day. Some trips are not taken simply because there is no time left to take them. This would suggest that a time budget approach to urban travel may very well yield far more interesting results than a gravity model approach. A recent survey of the literature in this area would suggest that it continues to have a high level of interest (Bhat and Koppelman, 1999).

#### *4.3. Sustainable transport*

Concerns over whether current transport systems are sustainable emerged in the late 1980s and early 1990s following the appearance of the Brundtland Report (World Commission on Environment and Development, 1987). A sustainable transport system is one that can satisfy current transport and mobility needs without compromising the ability of future generations to meet these needs. Researchers are currently of the viewpoint that our transport systems are not sustainable since they:

- use petroleum resources and this is a finite resource, which is rapidly diminishing based on some research;
- generate emissions that negatively impact air quality in urban areas;
- generate carbon dioxide and other greenhouse gases that are detrimental to the global environment in that they induce global warming;

- initially used coolants that destroy stratospheric ozone, and although an alternative is now available, this also damages stratospheric ozone;
- produce excessive numbers of incidents (formerly referred to as accidents), resulting in high numbers of fatalities and injuries;
- result in congestion that may border on gridlock.

The initial work in this area viewed sustainability as an environmental issue, and little attention was directed toward motor vehicle incidents or congestion (Transportation Research Board, 1997). However, a transport system that kills nearly half a million people a year globally should not be viewed as sustainable under any circumstances. Furthermore, if the system is congested to the point that flow ceases it should also not be viewed as sustainable (Black, 1996). Therefore, the notion has been broadened to include these two elements.

Critics argue that nothing is sustainable, and that is certainly true, but this conceptual approach can be used to structure nearly every type of research currently being undertaken in the field of transport geography, and most of transport generally, and it has value in that context even if we are in the final analysis unable to achieve a sustainable transport system.

Concerns over sustainability surfaced first in Europe, but a significant number of US transport researchers are also undertaking research within this structure. This paradigm has been assisted by a major transatlantic research undertaking supported by the European Science Foundation, the US National Science Foundation, and the EU. The initial effort in this area produced a series of workshops and an international conference on social change and sustainable transport, and several of the papers of that conference have now been published in a conference volume (Black and Nijkamp, 2002).

#### *4.4. Environmental justice*

Transportation projects often end up impacting different sectors of the population in different ways. Some of these impacts may be positive: the construction of a major arterial highway may reduce congestion and therefore travel time in getting to work for suburban dwellers. But this same project may produce negative impacts in that it may result in the destruction of neighborhoods or actually decrease local accessibility by the closure of streets and roads. Situations of this type should be caught by environmental impact assessments, which are intended to investigate all aspects of the environment, including the social and economic environment.

In the early 1990s it became apparent that some projects were being undertaken in the USA that did not require an environmental impact study but which might be creating unequal impacts on different segments of the population. President Clinton signed Executive Order 12898 in 1994 requiring each federal agency to

"make achieving environmental justice part of its mission by identifying and addressing, as appropriate, disproportionately high and adverse human health or environmental effects of its programs, policies, and activities on minority populations and low-income populations." Each agency issued its own environmental justice policy identifying how the order would be implemented.

Some of the early cases involve the locating of municipal incinerators or landfills. These are often located in poorer areas of a city because the land is cheap, and this may result in negative impacts on the poor or possibly a racial or ethnic group that may live in the area. Most transport researchers would say this results in disparate impacts, and the project should not be undertaken, but researchers do not have the final word in this matter – that rests with the courts.

The courts have said that even though the impacts are disparate and may even be discriminatory, it is necessary to demonstrate that the discrimination was intentional. This becomes a more difficult proposition. From a researcher's point of view the question is whether alternative sites were considered, and, if so, would any of these have produced fewer impacts. Obviously, the cost of the project would also be relevant and the number of people impacted.

However, since the signing of Executive Order 12898 there are now very few transport cases in the USA that have the unequal impacts described above.

There has not been a lot of transport geography research in this area, but it clearly is a transport research area where the skills of the geographer should be useful (Cambridge Systematics, 2002; URS Corporation et al., 2003). It is probably reasonable to conclude that most of the major cases of discrimination in the US transport sector occurred during the building of the interstate highway system and prior to the 1969 National Environmental Policy Act that required the preparation of comprehensive environmental impact statements. Nevertheless, there are undoubtedly situations where these impacts continue to be generated.

#### *4.5. Economic development*

There has always been a close relationship between transportation and economic development. It is a popular perception that investments in transportation will result in economic growth, but the reality may be quite different. First, we must clarify if we are talking about growth in developed or developing areas, and even then it is difficult to generalize what the impacts will be.

If a developed area has an advanced, uncongested transportation system and a high level of connectivity, this virtually ensures that it also has a high level of accessibility. Network additions (new links) to such a system will do little to stimulate economic growth. If the existing system has a high level of connectivity, but is congested, then it is possible that investments in the system will lead to positive economic benefits. This is essentially what Nadiri and Mamuneas (1996)

were able to demonstrate for the construction of the interstate highway system in the USA for the 1950s and 1960s. They were unable to demonstrate such benefits for the post-1970 period, but massive investments of the earlier type were not made in the post-1970 period.

Many transport geographers apparently have not looked at this research very closely. If they had they would have realized that the researchers ignored all externalities related to the building of the interstate highway system, e.g. additional environmental damage, and additional fatalities and injuries due to induced travel and higher speed travel. There should be little doubt that the project in question generated positive benefits, but these would have been significantly lower had the externalities been incorporated into the analysis.

One of the more interesting ideas to appear regarding economic development and transportation is the notion of the companion innovations of Garrison and Souleyrette (1996). This notion argues that transportation improvements may result in growth in other sectors, and this may lead to economic growth. Historically one can offer the expansion of the railroads in the USA and the companion innovation of the telegraph. This communication system allowed many other activities to flourish, which resulted in more growth than that due to the railroad alone. One can also envision a future transport system with some of the automatic guidance ideas currently being discussed in the intelligent transport systems area. Such a system would undoubtedly stimulate growth throughout the information technology area and generate significant economic growth beyond what would be stimulated by the transport system itself.

The potential positive, negative, and neutral impacts of transport projects in developing countries were identified several decades ago by Wilson (1966). It seems that many nations and funding agencies have missed the point. Major highways, such as the Trans-Amazon Highway, are constructed without the thorough *a priori* impact analysis they deserve. Leinbach (1995) has looked at highway development in South-East Asia. That research suggests that the continuing optimism regarding the benefits of highways and roads in developing areas cannot be sustained by the evidence. In particular, he notes that these projects actually do very little to affect mobility, and the rural poor see no benefits unless they have a transportation mode available for their use, or one is made available as part of the project, e.g. a rural bus service.

#### *4.6. Geographic information systems*

GIS are elaborate computer programs for handling geographically defined data. These include procedures for mapping points, lines, and areas, defined by their geographic coordinates, which are usually maintained in separate geographic databases. Associated with these geographic data are attribute databases that

include all types of information for the point, line, or area. For example, if the line is a highway, its attribute database might contain surface type, the number of lanes, grade, capacity, traffic volume, and so forth.

The mapping of geographic data using main-frame computers dates from the 1960s and 1970s. GIS that could be used on desktop microcomputers appeared in the 1980s. At that time they were interesting, but not very useful to the transport geographer. The primary problem was a lack of geographic databases. If one wanted to work with the highway network of Iowa, then that researcher would have to digitize the network of interest. This involved recording the geographic coordinates of every turn in the highway links of interest, and making sure that the systems being created were connected. This was an extremely time-consuming process that limited the use of these early systems.

This geographic database problem was solved in the USA with the passage of the Intermodal Surface Transportation Efficiency Act (ISTEA) of 1991. This act called for the creation of the US Bureau of Transportation Statistics (BTS), and that agency began to make available digital databases of all the major modal transport networks of the USA. One should not underestimate this research contribution by the BTS: it has enabled individuals and small research teams to carry out projects that would have been beyond their capability if the databases did not already exist.

A second problem with the adoption of GIS in transportation research was that many of the systems were really only elegant mapping systems. They performed no analysis or modeling. This problem has also been solved by vendors that now include nearly all of the commonly used models that one would generally encounter in transportation planning (e.g. Caliper Corporation, 2000). This would include the standard models for traffic generation, trip distribution, modal split, and traffic assignment. Many other procedures have been developed as part of some of these systems, and they have revolutionized computer use in this area (Thill, 2000; Miller and Shaw, 2001).

Critics of GIS use in transportation note that much of it is developed for aggregate data analysis, and it is not as easy to use with disaggregate data. There is probably some truth to this statement, although I do not know of anyone who has attempted to use the software in this manner. Vendors undoubtedly try to develop products for a market, and the transportation planning market tends to be the largest. There have been some interesting extensions to the use of GIS with individual movement as illustrated by the three-dimensional representations of personal travel by Kwan (2000).

#### 4.7. Network design

An area that has seen some significant development over the last couple of decades is the network design area, as typified by the work of O'Kelly and Miller

(1994). The primary focus of much of this work has been trying to identify algorithms for locating commercial aviation hubs for handling passengers. The objective is to have flights that are as close to capacity as possible. The work also has applicability to the overnight package delivery systems of companies such as Airborne, FedEx, and UPS (Kuby and Gray, 1993). Similar work has been undertaken for the movement of coal on the rail network of China by Kuby et al. (2001).

If one looks ahead there are indications that we will be seeing more and more small vehicles handling movement in urban neighborhoods. Since the mixing of large and small vehicles can be problematic, it might make some sense to develop procedures for identifying small vehicle sub-networks that keep these vehicles from interacting with large vehicles while also permitting the movement of such vehicles safely in the urban area. It is unknown if anyone is working on this problem.

## **5. Some concluding thoughts**

There are some areas not mentioned above that others might view as very important. One of these is technological change in the information and communication area. Included here would be the development of teleworking – a much better word than telecommuting, which is preferred in the USA – as well as the use of in-vehicle navigation systems, automated guidance systems, and similar developments that would generally be viewed as components of the intelligent transportation system (ITS) area. Teleworking may make some genuine contributions in the future, but at this point in time it posits far more questions than answers (Mokhtarian, 1997). It also appears to be decreasing as fewer employees and employers seem to have an interest in participating in such programs, at least in the USA. As for ITS, the area is full of promise (notably a major reduction in motor vehicle incidents and their attendant fatalities and injuries), but it is also apparent that these impacts have not been significant at this point in time.

## **6. Closure**

Transport geography is in many ways a developing field. Although this chapter is a review of recent progress, there has been an attempt to give some structure and organization to a field that is not well understood to those outside of it. The structure presented was also an attempt to focus future work. Some of the most impressive advances in geography over the past 50 years have been in transport geography; it also has the potential to do much more in the future.

## References

- Bhat, C.R. and F.S. Koppelman (1999) "A retrospective and prospective survey of time-use research," *Transportation*, 26:119–139.
- Black, W.R. (1996) "Sustainable transportation: a US perspective," *Journal of Transport Geography*, 3:151–159.
- Black, W.R. and P. Nijkamp, eds (2002) *Social change and sustainable transport*. Bloomington: Indiana University Press.
- Bureau of Public Roads (1965) *Calibrating and testing a gravity model for any size urban area*. Washington, DC: US Department of Commerce.
- Caliper Corporation (2000) *TransCAD: transportation GIS software, user's guide*. Newton: Caliper Corporation.
- Cambridge Systematics (2002) *Technical methods to support environmental justice issues*, NCHRP Report 8-36 (11). Washington, DC: Transportation Research Board.
- Garrison, W.L. and R.R. Souleyrette, II (1996) "Transportation, innovation and development: the companion innovation hypothesis," *Logistics and Transportation Review*, 32:5–38.
- Garrison, W.L. and J.D. Ward (2000) *Tomorrow's transportation: changing cities, economies, and lives*. Boston: Artech House.
- Goetz, A.R. (2001) "Deregulation, competition, and antitrust implications in the US airline industry. Fleming Lecture," in: *Association of American Geographers Annual Meeting*. New York.
- Haggett, P. and R. Chorley (1969) *Network analysis in geography*. New York: St. Martin's Press.
- Kuby, M.J. and R.G. Gray (1993) "The hub network design problem with stopovers and feeders: the case of Federal Express," *Transportation Research A*, 27:1–12.
- Kuby, M.J., Z. Xu and X. Xie (2001) "Railway network design with multiple product stages and time sequencing," *Journal of Geographical Systems*, 3:25–47.
- Kwan, M.-P. (2000) "Interactive geovisualization of activity-travel patterns using three-dimensional geographical information systems: a methodological exploration with a large data set," *Transportation Research C*, 8:185–203.
- Lalanne, L. (1863) "Essai d'une theorie des reseaux de chemin de fer, fondee sur l'observation des faits et sur les lois primordiales qui president au groupement des populations," *Comptes Rendus Hebdomadaires des Seances de l'Academie des Sciences*, 42:206–210.
- Leinbach, T.R. (1995) "Transport and third world development: review, issues, and prescriptions," *Transportation Research A*, 29:337–344.
- Miller, H.J. and S.-L. Shaw (2001) *Geographic information systems for transportation: principles and applications*. New York: Oxford University Press.
- Mokhtarian, P.L. (1997) "The transportation impacts of telecommuting: recent empirical findings," in: P. Stopher and M. Lee-Gosselin, eds, *Understanding travel behavior in an era of change*. New York: Pergamon Press.
- Nadiri, M.I. and T.P. Mamuneas (1996) "Contributions of highway capital and productivity growth," in: J. Madrick, ed., *Economic returns from transportation investments*. Lansdowne: ENO Foundation.
- O'Kelly, M.E. and H.J. Miller (1994) "The hub network design problem: a review and synthesis," *Journal of Transport Geography*, 2:31–40.
- Thill, J.-C., ed. (2000) *Geographic information systems in transportation research*. Oxford: Elsevier.
- Transportation Research Board (1997) *Toward a sustainable future: addressing the long-term effects of motor vehicle transportation on climate and ecology*, Special Report 251. Washington, DC: National Academy Press.
- URS Corporation and Public Policy Center, University of Iowa (2002) *Effective methods of environmental justice assessment*, NCHRP Report 8-41. Washington, DC: Transportation Research Board.
- Wellington, A.M. (1888) *The economic theory of the location of railways*. New York: Wiley.
- Wilson, A.G. (1967) "Statistical theory of spatial distribution models," *Transportation Research*, 1:253–269.
- Wilson, A.G. (1974) *Urban and regional models in geography and planning*. London: Wiley.

- Wilson, G.W. (1966) "Introduction," in: G.W. Wilson., B.R. Bergmann, L.V. Hinser and M.S. Klein, eds, *The impact of highway investment on development*. Washington, DC: Brookings Institution.
- World Commission on Environment and Development (1987) *Our common future*. Oxford: Oxford University Press.

*Chapter 3*

## INSTITUTIONS, LAND USE AND TRANSPORTATION

ROGER R. STOUGH

*George Mason University, Fairfax, VA*

### 1. Introduction

There are many factors that influence how land use decisions are made and the pattern or land use that evolves in a region or a nation. First and foremost is transportation cost. Most of the well-developed body of theory regarding land use is based on the minimization of transport cost among locations whether for residential or commercial use. While there is solid empirical support for most of the theoretical work, the reality of land use and land development patterns is one that is not always in line with theory. The reason is that institutions or social rules and rule structures also influence land use and transportation decision-making and thus land use patterns. Such institutions are driven by cultural and societal objectives, and are historical or evolutionary in nature, and thus are often not consistent with cost minimization principles. This is the reason that cost minimization theories or even historically based theories only partially explain land use decisions and patterns. The focus in this chapter is on examining and exploring the institutional factors that lie behind land use and transportation decision-making and the structure of the patterns that result from those decisions.

This chapter provides a framework for analyzing the institutions that underlie land use patterns and transport networks, and how these institutions influence decision-making in the context of urban and metropolitan development where most people live. Thus, primacy of institutions in the explanation of land use decision-making, transportation decision-making, and the spatial patterns that result from these decisions is assumed. At the same time it would be remiss, as indicated above, not to recognize the significant body of land use and transportation theory and supporting empirical research that exists. Thus the analysis begins with a review of the theoretical base. However, it then moves quickly on to explain that land use patterns are in part at variance with theory, and then to the institutional assessment of changes that have been occurring in urban and metropolitan land use and transportation patterns over the past two decades or so.

Major structural change in the form of rapid spatial decentralization of population and most other urban activities has been occurring in metropolitan land use patterns. This has resulted in significant consequences for transportation efficiency, operations and investment decision-making, and practice. Institutions and institutional analysis are important because theory does not fully explain why rapid decentralization and related changes in land use are occurring, and why these changes are stronger in the USA than in other parts of the world. A new typology for the institutional analysis of land use, adapted from the new institutional economics (Williamson, 2000), is introduced and used to structure the analysis. Before moving to the institutional analysis a summary of the theory that underpins urban land use and transportation analysis is presented along with recent trends in land use and transportation. Then the institutional analysis framework is presented as a prelude to the institutional analysis. Conclusions are offered in the final part of the chapter.

## **2. Theory and recent changes in land use and transportation**

The twentieth century witnessed the development and testing of a significant body of theory on the location of economic activity, i.e. land use, and transportation that included theories of urban form such as the concentric zone theory (Park et al., 1925), the sector or theory of axial development (Hoyt, 1939) and the multi-nuclei theory (Harris, 1943). Each of these provided sequentially more realistic models of the form of urban places, as expressed in land use patterns, and its relationship to transportation demand and infrastructure. Perhaps even more fundamental is agriculture land rent theory (Von Thünen, 1966) and its urban analog (Alonso, 1964). Rent theory argues that peripheral land generates less rent per spatial unit because of the higher transport costs required to move goods and people to and from an urban center. As a consequence, most cities have evolved around a hub and spoke transportation infrastructure, with higher rent-paying activities located nearer the urban core. There should be no surprise then that high-rise physical structures capable of paying considerable rents have tended historically to emerge at the urban geographic center.

It may seem ironic, given the argument presented in the classical and neoclassical rent literature, that the core-dominated model appears to be increasingly at odds with reality. Metropolitan regions, large and small, throughout the USA as well as increasingly in other parts of the world are rapidly expanding geographically as they become reorganized around multiple centers or edge cities (Garreau, 1991), the contemporary expression of the multi-nuclei concept. With this view the city is seen to expand spatially as nodes of activity emerge along major crossroads in the outer parts of urban areas. These peripheral activity centers tend to evolve in accordance with central place theory (Christaller, 1966),

where initially only low-level goods and services are provided. As growth continues, increasingly higher levels of goods and services are offered because of growing demand for them from positively reinforced expansion that spills over into the surrounding area from the agglomeration or concentration of activities (Gordon and Richardson, 1996). Today, most metropolitan areas in the USA (and increasingly in other parts of the world) are dominated by multiple centers where the historic core has become relatively less important and, in some cases, is no longer the dominant center. The newly emerging multi-centered metropolitan region requires significant institutional coordination and change to manage the nodal-driven decentralization that characterizes many metropolitan areas today. However, most metropolitan regions have limited experience and few organizations that are effective in a cross-jurisdictional metropolitan context in general and in the management of the rapid spread effects associated with the multi-nuclei metropolis.

## *2.1. Altered demand for urban transportation and land use*

The rise of the multi-centered metropolis has altered the historic pattern of demand for urban transportation as core-dominated hub and spoke systems of transport that were developed in the past are ill equipped to meet the growing demand between and among the new edge cities. The most rapid increase in demand for transport and related services for the past decade has been for cross-region travel. Metropolitan regions are having great difficulty coping with the associated increase in traffic congestion. This chapter illustrates the type of institutional factors that are influencing the development of metropolitan and corridor regions, how they are trying to manage the evolving multi-centered metropolis, and some of the lessons that are emerging. While much of the discussion draws heavily on the US experience, some comparative assessment is offered.

Transportation and land use are intimately linked, as settled land creates demand for transportation and vice versa. Thus, in general, urban places have more demand for transportation than rural ones. On the contrary, however, investment in transportation infrastructure may be the vehicle that is used to open up an area and make it more accessible, thus motivating land settlement and its productive use. The latter case is more often associated with earlier stages in the economic development process, where accessibility and linkage to markets is needed to enhance development and competitiveness of a region. In this chapter, concern is primarily with the former case, where development is driving transport demand. This focus is important because many major land use decisions and spatial growth patterns are driven by increases in transport demand. Today there is great dynamism and turbulence in the relationship between transportation and land use, as the spread effects of emergent and growing edge cities are changing

land use patterns and transport demand, and the perception of the economic and social value of land. The role of institutions in managing and influencing evolving relationships and solutions is fundamental to achieving acceptable and sustainable urban development outcomes at the local regional level as well as at the broader level of nations and multinational organizations.

## *2.2. Metropolitan decentralization*

Metropolitan decentralization has long been a characteristic of the US metropolis (Wardwell and Brown, 1980; Stanback, 1991; Nelson et al., 1995). However, a vast acceleration in the spreading out of metropolitan regions has occurred over the past two or so decades in the USA as well as to some extent in other countries such as Australia, countries in the EU, Japan, China, and India. The geographic area of developed parts of metropolitan areas in the USA has, in many cases, nearly doubled and in some cases nearly tripled during this period. For example, the federal definition of the Washington, DC, metropolitan region includes three state level jurisdictions (Maryland, Virginia and West Virginia) and a federal district – the District of Columbia – had 50% more counties in 2000 than in 1980, and grew from 80 000 ha to 1 300 000 ha in area. The Washington region is not atypical as metropolitan regions such as Atlanta Georgia, Dallas Texas, Houston Texas, Phoenix Arizona, Denver Colorado and a host of California metropolitan regions have experienced similar growth.

Metropolitan decentralization or sprawl has been driven by a diverse set of forces including the rise of the knowledge- and technology-intensive economy of the late twentieth century (Stough et al., 1998). Among the non-technology forces contributing to these rapid spread effects are factors such as lower land costs on the periphery, extensive highway systems lowering transportation costs to outer city locations, residential preferences of US citizens for the “marriage of town and country” living styles and the vision of a Jeffersonian rural lifestyle, deteriorating conditions in central cities, and, finally, a set of government policies that provide subsidies ranging from tax advantages to depreciation allowances to implicit subsidies in the form of building regulations and policies that discourage efforts to reuse older urban, and suburban, land (Ewing, 1994; US Office of Technology Assessment, 1995). Further, social issues related to spatial segregation by race and poverty may also be important factors (Bollens, 1988; Rusk, 1994). However, these forces have been present for many decades. So what has changed? The rapid development and ever-quickenning evolution of a new generic technology in the form of information and computer technology (ICT) is making a continuously changing and ever more spatially dispersed metropolitan economy and region not only possible but a reality (Niles, 1991; Kellerman, 1993; US Department of Transportation, 1993; Grantham and Nichols, 1994–1995).

Over the past decade or so the merging of computer and information technology has created vast increases in the ability of individuals and organizations to communicate. Local- and wide-area networks and the Internet coupled with new and emerging high-speed wire and wireless data transfer, and large capacity infrastructure hold the potential to connect all persons and organizations to each other. Illustrative of the impact of this trend and critical to understanding metropolitan spread effects is that an estimated 10 million or more people in the USA work from home or near home at least 1 day a week remotely or via telework. Similar patterns are emerging in the realms of “telelearning,” “teleshopping,” and “telebanking.” The cost of physically moving across metropolitan space is reduced for those who substitute virtual connections for commuter trips. For example, if the need for commuting to a central work place drops from 5 to 3 or 4 days a week due to remote working or telework, as it has for the 10 million or so teleworkers referenced above, living farther out from the work place in more rural residential settings, which to many represents a piece of the “American dream,” is attractive. In short, ICT creates a reduction in work trip friction and thereby contributes to spread effects (Stough et al., 2002). Other ways that ICT is contributing to reduction in the friction of metropolitan travel is via smart roads and intelligent transportation systems (ITS) that through information technology applied to the infrastructure and vehicles enhance the productivity or performance of existing infrastructure (Stough, 2001). In short, ICT is increasingly contributing to the substitution of communication for trip taking and more efficient movement, and thus metropolitan spread effects.

### *2.3. Edge cities*

However, all of these factors do not fully explain the structural nature and mechanics of the recent era of rapid decentralization. The evolution of edge cities is central. Edge cities emerged on the periphery of metropolitan areas in the 1980s and 1990s. They appeared at or near the confluence of major transportation infrastructure networks and near former bedroom communities that housed white-collar and highly educated workers who commuted to work in the center city. Such attributes were important for new emergent and often technology-intensive enterprises because of a need for access to other infrastructure, e.g. airports and business services, to high-quality workers such as those found in the bedroom communities on the edge, and to relatively inexpensive land that could only be found in peripheral locations. As a consequence of the role and importance of edge cities in the rapid decentralization of metropolitan areas, they are briefly described in the following paragraphs.

Following Garreau (1991), edge cities are large and diverse knowledge age urban concentrations that have appeared on the urban landscape over the past two decades or so and have become the standard form of US urban place. Edge

cities, in many ways, play the same role as the old traditional downtown centers did regarding jobs, shopping, entertainment, services, and housing. However, the new cities on the edge are often sprawling, seem erratic, are peculiar to look at, and often spill over political boundaries.

Edge cities are employment centers that have several attributes (Garreau, 1991). They:

- cater primarily to commercial office buildings (the workplaces of the knowledge age);
- contain most of the commercial office and retail development that occurred during the various stages of urban growth experienced during the past three decades;
- have a population base that is dominated by white-collar workers;
- offer a variety of goods and services as well as entertainment and restaurants;
- are perceived as one place, and an end destination for mixed use no matter how sprawling they may be;
- rarely have formal political government with elected political officials.

Edge cities have at least 25 000 jobs and several million square meters of commercial office space, at least 100 000 square meters of retail space, a dominant middle class and highly educated population, and are home to many technology intensive companies (Stough et al., 1998). Edge cities are an expression of one of the most profound changes in land use experienced in the recent past. They are where most new businesses have located over the past 20 years and have thus become new urban-industrial complexes in areas that a few years ago were located on or beyond the margins of metropolitan development. Edge cities have created highly "urbanized counties in proximity to, but increasingly functionally independent from, the central cities of many metropolitan regions" (Stough et al., 1998). Edge cities are important because they represent a new urban form that has been pivotal in the changing land use reality of the USA. As such, it has been central to the sprawling pattern of land use that has arisen in the US city and thus to increased tension between land use development, policy, and transportation investment patterns.

Edge cities played a major role in the vast decentralization of economic and residential activity that occurred in the past two decades or so in US metropolitan regions as well as in other countries. No examination of metropolitan decentralization in the late twentieth century should be made without considering the pivotal role that edge cities have and continue to play.

### 3. Institutions

Analysis of land use decision-making has mostly utilized case study or interpretive methodology, although in an interesting break with this tradition Wang et al.

(1998) tested several land use decision-making hypotheses derived from social judgment theory using an experimental design methodology with human subjects. Nonetheless, much of the published research regarding land use decision-making fails to bring institutions formally into the analysis. After cost minimization theory, institutions are probably the most important factor in explaining land use and transport decisions and patterns of land use. In an effort to provide some structure for the institutional analysis in this chapter, a typology developed by Williamson (2000), a disciple of the new institutional economics, has been adopted and used here. Thus, institutions are defined from the perspective of the new institutional economics. Below, Williamson's typology is outlined and illustrated and followed in Section 4 with an application of the typology to land use decision-making.

The new institutional economists, after North (1990), define institutions as rules and more specifically as “the rules of the game ... or alternatively ... the humanly devised constraints that shape human interactions in a society” (Clingermayer and Feiock, 2001). As such, institutions serve to guide individual behavior, reduce uncertainty, and stabilize public choices that would otherwise be even more turbulent than they are. In this sense, institutions produce more predictable and trustful environments or ones that are more institutionally thick (Amin and Thrift, 1995). Places that are institutionally thick tend to reduce decision transaction costs more than places with fewer institutions, less inter-institution or organizational interaction, and fewer informal conventions, habits, and routines. It is important to note, however, that institutional thickness can also become too cumbersome, leading to diminishing returns. Places that have many strong and thus coveted traditions, and heavy regulatory over-burdens and are thus rule bound, as observed in some regions that have attempted to use strong land use regulation to control growth, have found their institutions a hindrance to forming and implementing strong regional development and sustainability policies as well as more social-oriented policies such as diversity.

There is a difference between organizations and institutions. It is important to note this distinction because in everyday speech and language these are often used synonymously. Organizations are groups with specific and common purposes and are created and operated in a context defined by the institutions or rules of society. Organizations, like institutions, also play a role in guiding individual and collective action. As North (1990) notes, “organizations are the players and institutions are the rules.” So, for example, the zoning board, land use planning office, and local land use authority are organizations, and the rules, statutes and administrative orders, and board decisions (both formal and informal) that create them and frame their operations are the institutions.

Economic and societal performance and behavior are fundamentally influenced and often guided by institutions and the way they develop. In this sense, they help to manage and reduce market imperfections such as cheating, corruption, and

lack of transparency, inefficiency, free riding, transaction costs, and tragedy of the commons. Thus, institutions determine the long-run performance level of the economy and, in turn, society. Institutions reduce uncertainty by defining the margin at which organizations operate, thus making the rules of the game and agent behavior predictable (North, 1990). Institutions are important at the interface of transport and land use because they are the formal and informal or implicit rules by which such decisions are made. Thus, the delicate relationship between land use and transportation is guided, maintained, and/or made turbulent by the institutions or rules.

The new institutional economists provide a framework for analyzing institutions and the ways in which they may impact decisions and behavior, both positively and negatively. Williamson (2000) views institutions as taking one of four forms: informal, formal, governance, and resource allocation/employment related. Examples of informal institutions are deeply embedded values, norms, customs, and traditions. These are powerful conditioners of behavior but change only very slowly. However, when change does occur in an informal institution, behavior changes rapidly and profoundly, e.g. the terrorist events of September 11, 2001 have significantly impacted privacy and accessibility norms in the USA and other parts of the world. This is a case where an extreme event impacted informal institutions almost immediately. However, one could argue that this change was brought on only over a long period of 20–30 years, coinciding with the gradual growth and spread of terrorist actions in general.

Formal institutions are Williamson's second category. These rules are codified as laws, regulations, and administrative orders, and include, for example, such things as property rights, judicial orders, and administrative statutes. Formal institutions change more quickly than informal ones, but still over fairly long periods such as decades unless there are radical changes in the environment within which the rules or institutions apply. For example, civil and water rights legislation in the USA as well as in other countries has involved decades of debate and multiple trial and error efforts at legislation to at best produce modest incremental change. Efforts to alter land use regulations to address so-called inefficiencies of sprawl and to support smart growth objectives have been underway for more than a decade with modest results.

Williamson's third institutional level is governance. Here, institutional change occurs relatively fast, often measured in years as opposed to decades. Governance institutions are rules (minor laws, administrative orders, regulations, and policy directives) that are used to change how government and organizations that involve governance such as planning and zoning boards conduct business and transactions with other actors and agents.

Finally, at the fourth level are the action and behavior patterns of a diversity of actors in the decision environment ranging from government agencies to firms and to non-profit associations, e.g. neighborhood organizations. Institutions at

this level are about allocating resources directly related to near-term productivity (individual and organization level) and operational outcomes. These institutions are changing continuously because they have highly distributed and varied consequences. However, the consequences at the societal level are quite small and often relatively insignificant in terms of long-run outcomes. They involve decisions and actions about production, delivery, and resource acquisition and use, and process and occur in a context measured in days, weeks, and months. Making decisions about a zoning variance request or changing a minor land use plan implementation component are examples of the fourth level of institution.

Williamson's typology (2000) provides a framework for examining policy arenas as well as relationships between different arenas and, thus, a way to identify and understand the forces that are guiding action and behavior in specific contexts. It also provides a way to identify efficiency, effectiveness, and equity problems and policy intervention strategies. In the following section this institutional framework will be used to examine land use sprawl in the USA.,

#### **4. Institutional analysis decentralization in metropolitan regions**

When flying over Germany, France, or Portugal, for example, on a clear day one observes a pattern of urbanization that sharply demarcates urban from rural land use. Village and even city boundaries have clear breaks between urban and rural land usage. But land use patterns observed while flying over the USA reveals at best a fuzzy demarcation between urban and rural uses. There, metropolitan areas are vast and exhibit a gradual but intermittent shift from urban to rural use of the land that shows the consequence of the decentralization or sprawl described above. Why are the patterns so different? The answer lies largely in the different institutions and institutional barriers that guide the land use decision process. At the deepest and most profound level, US values and culture or informal institutions are responsible for the pattern of land uses found there.

##### *4.1. First-level institutions: US values and culture and metropolitan land use patterns*

American exceptionalism (Lipset, 1990, 1996), erected on the nineteenth century observations of de Tocqueville, provides some insight into why decentralization and sprawl are so characteristic of the typical land use patterns found in the USA. The settling of the USA took great individual effort that led to a strong individualist tradition, which in turn became a defining element of US culture. While frontier life spawned a kind of independence and thus individualism not

found in most parts of Europe, where most early US settlers came from, it also evolved elements of cooperation but at the local community level rather than through government organizations. Cooperation to solve governance or problems of the people was not based primarily in the state or dependent on leadership from the federal state. Such leadership was highly decentralized, e.g. at the level of dispersed farmsteads and villages that early on defined the communitarian values that also typify US culture. Later in US history this form of cooperation evolved into and became embedded in voluntary associations that often assumed responsibility for community action, thus adopting various roles that government plays in most other countries. This is what de Tocqueville was so taken with during his extended visit to the USA. Individualism and community level self-reliance via voluntary association are two of the major attributes of the American exceptionalist view. This community approach is also seen in the tradition of philanthropy in the USA that often provides the resources for non-government community action. Even more interesting is the view that philanthropy is the primary vehicle for redistributing the vast accumulation of wealth by entrepreneurs that is permitted in the USA (Acs, 2003).

The founders of the USA also created a country with weak political institutions. This, some argue, is because the country was created on the platform of a rebellion and then a revolution against what was perceived as the excessive use and abuse of power by a sovereign king. This is symbolized by the multiple checks and balances and separation of power between the federal state and the individual states that characterize the US government system. These institutions or rules were consciously created by the founders and signers of the US constitution to ensure that government power would never be so highly concentrated in the USA and would require considerable debate among representatives of different and conflicting views and stakeholders before major decisions could be made.

Lipset and others who adopt the exceptionalist view see the USA as a different type of government and as having a different governance form. Its citizens or people are more individualist oriented and more inclined to solve and attack problems through non-government organizations. Further, these values are reinforced by virtue of the fact that the role of government is relatively weak compared with most other countries. The laws, statutes, rulings, and decisions that lead to the sprawling land use patterns that characterize US urban settlement are in large measure due to this relatively unique set of values.

#### *4.2. Second-level institutions: formal institutions*

Formal institutions are rules that are established in statutes, regulations, or other public policy actions such as administrative orders and that are enforceable by law. Land use regulations and statutes in the USA are largely established and

implemented at the local jurisdictional level (counties, cities, and towns) despite the fact that, as many other countries have found, special interests may secure advantageous land use positions and decisions regarding their specific interests. To some extent, where cross-state and national level interests are at stake, the federal state and the state within which the local jurisdiction is located require certain land use and zoning plan components such as practices required to meet environmental maintenance requirements, e.g. wetlands protection and the need for environmental impact analyses where transport infrastructure projects are of a scale that will have potentially large negative spillover effects. However, the invasiveness of these interventions from higher levels of government is much less than in most, if not all, European Union countries, Australia and other former colonial states. The role of sub-national states in land use planning and zoning in the USA is also relatively modest, as most have ceded considerable latitude to the jurisdictions. For example, there is no sub-national state in the USA that has retained and maintained responsibility for land use planning and zoning regulation, and enforcement.

One way to understand the relatively modest role the federal state and the states have played in land use and zoning regulatory, and statutory action, i.e. second-level institutions, is to view it in the light of first-level institutions. Decentralization of formal land use institutions is consistent with a culture or society such as the USA where strong individualist leanings and a strong distrust of government are dominant values. Further, it is consistent with a view that values control action that is managed as close as possible to the sphere of action, in this case the “community” level, which is roughly synonymous with the jurisdiction. Even in cases such as Oregon, where states have intervened and assumed a stronger role in managing the development and use of land, it has not worked particularly well, mostly because of the powerful underlying first-level institutions that continue to be maintained in the form of the values of individualism, distrust of government, and non-government action directed at addressing and achieving public and social goals.

Over a decade ago Oregon adopted a state statute that placed a cordon beyond the development perimeter of Portland, the largest city in the state. The purpose of this statutory action, or institution, was to force development inside the cordon and, therefore, to achieve more efficient land development and, by so confining development, reduce the local potential for sprawling urban landscapes such as those evolving in and around other US metropolitan regions. Development and the possibility for development were thus to be highly constrained by law. However, a few years after implementation, growth, instead of being confined inside the controlled zone, began to bypass the cordon to smaller cities and urban concentrations well beyond the buffer, where development was not constrained by this law. The result has been decentralization that leapfrogged to satellite cities beyond the perimeter rather than incremental sprawl driven by

edge city development. As such, decentralization and loosely distributed urban development or sprawl found in most US metropolitan areas has also occurred in Portland.

#### *4.3. Third-level institutions: governance institutions*

Governance institutions are rules that include minor laws, administrative orders, regulations, policy directives, and decisions mostly focused at the local level in the USA. While the cordon statute in Portland, Oregon, has been treated as a governance institution and was enacted at the state level, it is an exception in the US context. Moreover, representatives from the Portland region lobbied the state to enact it. The difference between formal institutions and governance institutions here is somewhat artificial but may be viewed as the difference between the more durable and slow-changing rules and regulations of the federal state and sub-national state that frame the context within which governance institutions operate, compared with more ephemeral local level rules and laws such as development and building requirements, and the rules that define how variances to land use and zoning regulations may be made. Governance institutions tend to reflect the preferences of local stakeholders and the decisions that come out of the averaging of preferences that frequently occurs in such decision-making. Further, they tend to be less durable than the formal institutions of the state. For example, land use and zoning regulations have frequently been made more restrictive in or near the end of a robust economic period, where negative spillover effects of growth on the landscape are manifested in higher levels of traffic congestion and lagging provision of schools and police and fire services. However, within a few years as recessionary effects are experienced with increased unemployment and other negative economic effects, and as service provision begins to catch up, liberalization of governance institutions may occur as the need for growth becomes a stronger locally perceived need. This sort of change in the rules illustrates the third-level or governance institutions.

Two cases from the Washington, DC, metropolitan region illustrate how local governance institutions respond to relatively short-term changes in the decision environment. The Washington region has been one of the fastest growing metropolitan regions in the USA for the past 30 years. Consequently, extensive edge city development and sprawl has occurred on its urban periphery. Two counties on that periphery, Prince William and Loudoun, absorbed large amounts of this growth. Being located on the periphery meant that land was available and at relatively low cost, but the average land values were even less in Prince William County, meaning that young workers in the family formation and income-constrained stage of household development were most attracted to residential sites in that county. Land values in Loudoun County were somewhat higher

because a large international airport is located there, along with better transport connectivity to high-development areas located nearby but closer to the urban core, e.g. Tysons Corner.

In the early and mid-1990s both counties experienced exceptionally high population growth rates but Prince William County, with its poorer transport and economic links to the existing economy of the region, did not experience significant edge city development, and, consequently, job growth and economic development lagged there. As demand for services rose rapidly from the tens of thousands of new and young families that settled in the area, residential taxes increased rapidly, and in turn began to offset the attraction of low land values. Considerable effort was put into economic development initiatives and related commercial development that dominated local initiatives in an effort to increase the commercial tax base and thus shift the tax burden in part from residents to commercial enterprises. The strategy worked well, and today Prince William County is shifting attention to attending more carefully to the land use and zoning regulatory framework.

Loudoun County on the contrary not only experienced rapid population and residential growth but also had a more even experience with commercial development. Consequently, much more attention early on was focused on the land use issues and the regulatory or institutional framework. In the late 1990s Loudoun County officials and leaders struck upon a plan that would encourage development in the more centrally located half of the county and would move to exclude further development from the more peripheral part of the county. Land use regulations have recently been passed to this effect, although it is unclear how the plan will work in the face of continued growth (Milligan, 2003).

What is important about governance institutions in the USA is that they are for the most part phenomena of local jurisdictions and they are less durable than the formal rules and regulations set forth at higher levels of government. But these more formal governance rules are general and relatively modest in scope. The two examples from the Washington metropolitan region show the primacy of the local jurisdiction in setting up and changing governance institutions and their relatively dynamic nature.

#### *4.4. Fourth-level institutions: resource allocation and short-term outcomes*

One trend in the USA as well as other countries over the past 30 years is the increased breadth of participation of stakeholders in land use and transport decision-making at all levels of policy and institutional formation. Today, there is a large diversity of actors in the decision process at even the most fine-grained local level. Such actors or stakeholders include not just professionals, government officials and elected officials, but industry groups, technical groups, developers,

safety representatives, and environment and heritage representatives, to name just a few. Institutions and rules that guide everyday operations and performance in land use development, and transport investment and construction policy, have become more rigid because selected stakeholders increasingly used operational flexibility as a vehicle to press their own interest and to achieve desired outcomes. As a consequence, decisions that were considered to be made by professionals and experts at an earlier time have increasingly become more rule and regulation bound, thereby increasing the time required to make resource allocation and even minor implementation decisions. In short, what in many cases may have been a fourth-level institution a decade or two ago has in practice become a third-level institution today. This is due to the increasing complexity of the stakeholders in most land use and transport decisions and, of course, the increasing complexity of the decision context.

Further evidence of this trend is illustrated by the development and legitimization of “private communities … governed by homeowner’s associations and guided by detailed rules of governance that are more or less equivalent to the administration of neighborhood zoning” (Deng et al., 2002). These common interest developments (CIDs) represented only about 1% of the residential stock in the USA in 1970 but now represent about 15%. CIDs and the rules they administer influence considerably where developments occur and how land is used in urban areas and as such add another regulatory dimension to land use decision-making and indirectly to transportation decisions and management.

#### *4.5. Institutional analysis: conclusions*

The above discussion presents a definition of institutions and then uses it to establish a typology for institutional analysis. The typology has four levels of institutions. At the most fundamental level are informal rules that manifest themselves as values, culture, habits, and traits characteristic of a group or society. These powerful but slow-changing institutions are like the stage upon which land use policy and transport decision-making occurs. The way in which these institutions influence the nature of formal institutions that guide land use development at the federal state and state levels in the USA was illustrated and shown to be responsible for the delegation of most such decisions to the local level. In fact the rise of importance of CIDs may in part be a consequence of the individualist and anti-statist values that are so deeply ingrained in the US ethos. This is important because these institutions help to explain why formal institutions at higher levels of government in the USA are so weak compared with other countries. Formal institutions also exist at the local level, the third institutional level, but are less durable and more subject to the whims of events and changing conditions. This was illustrated with the case analysis from the Washington, DC, metropolitan region in

the USA. Finally, fourth-level institutions, those that guide resource allocation and professional implementation and requirements decisions, were found to be increasingly more codified as a consequence of the broadening and deepening of stakeholder participation in US land use and transport decision-making.

## 5. Conclusions

The purpose of this chapter was to explain recent trends in US land use development and transportation and to examine institutions as explanatory factors of those trends. It was argued that land use and transportation actions mirror each other, with growth and new technology driving the need for expanded use of urban land and in turn transportation infrastructure (or at the minimum driving the need for increased productivity in the use of transport infrastructure). The sprawling trend in urban land development was explained by identifying the factors driving this process. Among these forces, the rules or institutions that guide how land is developed and transport infrastructure decisions are made were considered to be critical in explaining why the spotty, fuzzy, and often unconnected but always expanding urban development of the USA is far less organized than in many other Western countries. A four-level institutional typology was developed and used to examine this difference.

Informal institutions or the first level were seen to be the most powerful and enduring force, and the strongest explanatory factor. The exceptionalist (Lipset, 1990, 1996) view of the US core values of individualism and weak government institutions evolved out of a distrust of government coupled with a community orientation toward the use of non-government organizations to accomplish social goals form the first-level institutions that most help to explain the USA's failure to achieve tighter, more compact urban land use patterns. They also may be viewed as a major reason why the adoption of so-called "smart growth" practices have been relatively slow, especially compared with other countries. This result is instructive in that it suggests that understanding significant divergences in land use and transport decision-making across countries is either a result of quite different institutions that begins with first-level informal institutions or represents perhaps a fairly high level of innovativeness on the one hand or ignorance on the other.

## References

- Acs, Z.J. (2003) *Entrepreneurial capitalism: if America leads will Europe follow?* Working paper. Baltimore: MD: Merrick School of Business, University of Baltimore.  
Alonso, W. (1964) *Location and land use: toward a general theory of land rent*. Cambridge, MA: Harvard University Press.

- Amin, A. and N. Thrift (1995) "Globalization, inatitutional 'thickness' and the local economy," in: P. Healey, S. Cameron, S. Davaoudi, S. Graham and A. Madani-Pour, eds, *Managing cities: a new urban context*. New York: Wiley.
- Bollens, S.A. (1988) "Municipal decline and inequality in American suburban rings: 1960–1980," *Regional Studies*, 22:277–285.
- Christaller, W. (1966) *Central places in Southern Germany* (C.W. Baskin, translator.) Englewood Cliffs: Prentice-Hall.
- Clingermayer, J.C. and R.C. Feiock (2001) *Institutional constraints and policy choice: an exploration of local government*. Buffalo: State University of New York.
- Deng, F.F., P. Gordon and H.W. Richardson (2002) *Private communities, market institutions, and planning*, Working paper. Los Angeles: University of Southern California.
- Ewing, R. (1994) "Characteristics, causes, and effects of sprawl: a literature review," *Environmental and Urban Issues*, winter:1–15.
- Garreau, J. (1991) *Edge city: life on the new frontier*. New York: Doubleday.
- Gordon, P. and H.W. Richardson (1996) "Beyond polycentricity: the dispersed metropolis, Los Angeles 1970–1990," *Journal of the American Planning Association*, 62:289–295.
- Grantham, C.E. and L.D. Nichols (1994–1995) "Distributed work: learning to manage at a distance," *Public Manager*, winter:31–34.
- Harris, C.D. (1943) "A functional classification of cities in the United States," *Geographical Review*, 33:85–99.
- Hoyt, H. (1939) *The structure and growth of residential neighborhoods in American cities*. Washington, DC: Government Printing Office.
- Kellerman, A. (1993) *Telecommunications and geography*. New York: Halsted Press.
- Lipset, S.M. (1990) *Continental divide: the values and institutions of the United States and Canada*. London: Routledge.
- Lipset, S.M. (1996) *American exceptionalism: a double-edged sword*. New York: Norton.
- Milligan, J. (2003) "Showdown in Loudoun," *Virginia Business*, 18:4.
- Nelson, A.C., W.J. Drummond and D.S. Sawicki (1995) "Exurban industrialization: implications for economic development policy," *Economic Development Quarterly*, 9:119–133.
- Niles, J.M. (1991) "Telecommuting and urban sprawl," *Transportation*, 18:411–432.
- North, D. (1990) *Institutions, institutional change and economic performance*. Cambridge: Cambridge University Press.
- Park, R.E., E.W. Burgess and R.D. McKenzie (1925) *The city*. Chicago: University of Chicago Press.
- Rusk, D. (1994) *Cities without suburbs*. Baltimore: Johns Hopkins University Press.
- Stanback, T. (1991) *New suburbanization: challenge to the inner city*. Boulder: Westview Press.
- Stough, R.R. (2001) *Intelligent transportation systems: cases and policies*. Cheltenham: Edward Elgar.
- Stough, R.R., K.E. Haynes and H.S. Campbell, Jr (1998) "Small business entrepreneurship in the high technology services sector: an assessment for edge cities of the US national capital region," *Small Business Economics*, 10:61–74.
- Stough, R.R., K. Button and P. Nijkamp, eds (2002) *Transport and information systems*. Cheltenham: Edward Elgar.
- US Department of Transportation (1993) *Transportation implications of telecommuting*. Washington, DC: US Department of Transportation.
- US Office of Technology Assessment (1995) *The technological reshaping of metropolitan America*, OTA-ETI-643. Washington, DC: US Government Printing Office.
- Von Thünen, J.H. (1966) *Von Thunen's isolated state* (C.J. Friedrich, translator). Chicago: University of Chicago Press.
- Wang, M.S., J.K. Fang and W.M. Bowen (1998) "A multi-criteria experimental comparison of three multiple attributes weight measurement methods," *Journal of Multiple-Criteria Decision Analysis*, 7:340–350.
- Wardwell, J.M. and D.L. Brown (1980) "Population redistribution in the United States during the 1970s," in: D.L. Brown and J.M. Wardwell, eds, *New directions in urban–rural migration*. New York: Academic Press.
- Williamson, O.E. (2000) "The new institutional economics: taking stock, looking ahead," *Journal of Economic Literature*, 38:598–613.

*Chapter 4*

## TRANSPORTATION LOCATION AND ENVIRONMENTAL JUSTICE: A US PERSPECTIVE

KINGSLEY E. HAYNES

*George Mason University, Fairfax, VA*

### 1. Introduction

In the USA, research on environmental justice has blossomed in the last two decades. Environmental justice has become a nationwide social and, increasingly, populist movement centered on understanding and reducing the uneven and disproportionate impacts of environmental hazards on minority and poor populations in urban and rural regions. With the initiation of various federal programs, federal policy-making in the domains of housing, the environment, energy, commerce, land use, and transportation has also begun to focus on equity, health, and other impacts of environmental pollutants and the location of toxic and high-pollution-generating facilities.

In the realm of transportation, environmental justice means that transportation investments should be studied carefully to determine the nature of probable impacts, both favorable and adverse. During the past four decades, massive investments have been made in transportation infrastructure. One of the largest examples of these investments has been the US interstate highway system, completed in the late 1980s and early 1990s. As time has passed, however, it has become increasingly clear that not everyone has benefitted from these investments; evidence with regard to adverse impacts is emerging. Some populations, often low-income or minority, have been adversely affected by the construction and operation of these facilities through the increase in noise and air pollution levels and a negative impact on housing values. However, the state of knowledge in this area is subject to the constraints of scientific modeling. These constraints include assumptions used in modeling and tracing the pathways of pollutants, in characterizing local meteorological conditions, and in assessing the toxicological effects of waste sites on human populations (Greenberg, 1993). From a legal perspective, issues of due process, intent, and discrimination in siting are prominent. As a result, there exists some confusion with regard to the proper

inferences to be drawn from such models for policy-making and for promulgation of laws and rules to minimize the negative consequences on health and other human activities (Hird, 1993).

Bryant (1995) describes environmental racism as the deliberate targeting of minority groups for siting of undesirable land uses, which lead to disproportionate hazard exposures for those groups. From a general perspective it is argued that environmental justice requires that everyone has access to safe and clean neighborhoods, adequate jobs, quality schools, and "sustainable communities." Environmental equity as defined by Pollock and Vittas (1995) refers to the "extent to which physical and economic burdens of environmental disamenities are evenly distributed across society."

Location of facilities and population is not solely a physical process. Rather, it is a result of broader socio-economic and political processes. Hence, identifying points of intervention to minimizing impacts must be based on an understanding of both social and physical factors influencing environmental inequities.

The extent of any intervention must be based on a reasonable and systematic evaluation of the inherent trade-offs in social, economic, political, legal, and scientific aspects of environmental needs and concerns. For example, transportation planning has long considered the jobs-housing match as an important land use balancing consideration for communities and individuals in urban areas (Stutz, 1986). Environmental justice must be addressed with similar specification of location-impacts trade-offs. Similarly, community input and concerns must also be incorporated in any planning and policy effort in order to maximize benefits and minimize costs, including negative externalities.

The majority of the existing literature on environmental justice takes an indirect approach to this issue. Most studies limit their examination to the geographical coincidence of facilities, releases, and poor and minority populations using census of population and housing data. Such research is valuable in addressing environmental inequity concerns, but it is limited with regard to presenting conclusive evidence on the pathways through which the concentration and distribution of environmental and other inequities occur. This is especially true in terms of appreciating the disproportionate impacts on poor and minority populations. One of the more promising areas for research is to use interdisciplinary and systematic approaches to understand and explain the concentration and location of facilities and poor and minority populations resulting from the nexus of transportation, land use, housing, labor, business, and individual decisions. Social and economic impacts cannot be considered as separate and independent from other broader processes (Bowen and Haynes, 2000a). Further, public intervention for market failures requires comprehensive accounting of social and economic costs and benefits from facility location for communities and responsible jurisdictions (Cutter, 1995).

This chapter examines the impacts of transportation on toxics release inventory (TRI) facility location in the USA and the relationships between the location and other socio-economic elements of the urban landscape in an assessment of spatial inequities. The analyses are based on extending previous research in the Cleveland metropolitan area of the USA (Bowen et al., 1995). The new research question is whether transportation infrastructure, particularly highway and road access, influences the siting, location and use decisions of noxious facilities. Systematic considerations of such locational decisions is lacking in current social science research. Although there is a considerable literature on the role of transportation in the location of firms, housing, retailing, and a variety of other land uses as well as a literature on risk assessment of natural and man-made hazards, the link between these literatures is quite limited. The next section reviews the research on environmental justice concerned with the geographical coincidence of TRI facilities, releases, and poor and minority populations. The third section discusses a small but growing literature on transportation and environmental justice. The fourth section presents the findings of our analysis with respect to Cleveland, and the final section concludes the chapter with some generalizations and hypotheses for future research.

## 2. TRI geography and environmental justice

### 2.1. *Environmental justice and TRI*

In recent decades, research examining environmental justice and environmental equity issues has accelerated. A major starting point of empirical studies in the USA was the establishment of the TRI/right to know legislation that was authorized under the amendments to the 1986 Superfund Amendments and Reauthorization Act (SARA). The Emergency Planning and Community Right-to-Know Act (EPCRA) of 1986 focused on assisting communities to prepare for the possibility of chemical spills and other emergencies. In 1987, the first set of data under the TRI provision was released. While largely a reporting function, the use and analysis of these data both by environmental and community groups and by industry itself assisted in efforts to document the potential and real impacts of toxic chemicals (Taylor, 2000).

Community groups in the USA began actively using TRI data to document the types of air, water, and land releases in different parts of the country (Kunreuther and Easterling, 1996). The now famous United Church of Christ (UCC) report released in 1987 coined the term “environmental racism.” The environmental justice movement developed and evolved by highlighting the importance of race and income in environmental-justice-related issues (Commission for Racial Justice, 1987). The movement was further strengthened with Presidential Executive

Order 12898 in 1994, requiring all federal agencies to develop an annual plan "that identifies and addresses disproportionately high and adverse human health or environmental effects of its programs, policies, and activities." In 1995 the US Environmental Protection Agency (EPA) added 325 more chemicals to the TRI list. More recently, some individual states (e.g. New Jersey and Massachusetts) have enacted stricter laws, requiring manufacturers and users of toxic chemicals to disclose toxics use and to target source reduction efforts. These toxic reduction and use requirements effectively highlight the importance of pollution prevention at its source (Bryant, 1997).

While as a popular movement the environmental justice movement must be considered as relatively successful, the complexity of the issues in addressing its final impacts on communities and neighborhoods, particularly environmental equity or inequity, has been more problematic. There exist a multiplicity of methods and often contradictory findings on the disproportionate impact of environmental hazards on low-income and minority groups (Bullard, 1994). Differences between studies in terms of geographic scales and level of analyses (Bowen and Haynes, 2000b) include the following: (1) whether to use zip codes, census blocks, or counties; (2) measurement of both socio-economic conditions and TRI impacts (for the former, what variables to use to represent wealth, income, housing conditions, and other socio-economic profiles; what constitutes appropriate control or comparison groups, whether nationally or regionally specific impacts are to be assessed; and, for the latter, whether site locations, quantity of releases, types of releases, or release toxicity are the appropriate impacts to examine); (3) the accuracy and limitations of data that have been questioned; and (4) which theoretical drivers appear to be important (racism, ignorance, markets, conflict theory, etc.) and to what degree (Bowen, 2001).

This popular movement on behalf of environmental justice found support in federal law, namely the mandate of the US Civil Rights Act of 1964. This act requires federal programs to be non-discriminatory, and thus it encompasses federal environment protection activities. Although research on toxic, hazardous, and commercial waste had long-documented inequalities in the siting of waste facilities (Collins, 1992), the federal court first considered these inequalities as a point of law in a case involving the disproportionate placement of landfills among black residents near Richmond, Virginia.<sup>a</sup> Although illegal discrimination was not proven, owing to "the intent issue" (i.e., the plaintiff's necessity to demonstrate "racial animus" or racist intent in order to sustain a charge of discrimination), this case clarified the terms of the legal debate. In cases of environmental equity,

<sup>a</sup>*Residents Involved in Saving the Environment v. Kay* (1998). 768F. Suppl. 1141 (ED Va. 1998) (RISE I); 768F. Suppl. 1144 (ED Va. 1991) (RISE II).

unlike other cases of discrimination, the federal courts have not resolved the evidentiary presumption that the totality of circumstances can be used to prove intent in what appears to be cases of racially based environmental discrimination (Bowen et al. 1995).

Measurement issues also are important with regard to the impact of environmental inequities if and when they are established. Due to the lack of knowledge on the toxicity of a majority of chemicals, different methods have been developed for chemical ranking and scoring schemes. At this time, many of these schemes have been built on different assumptions and consequently differ in their final ranking and scoring as to TRI chemicals. A number of these methods identify exposure potential and hazardous locations but cannot be used to do actual risk assessments. Risk assessments require more facility and specific environmental information that would relate chemical releases and their impacts on the surrounding areas as well as account for cumulative impact of multiple chemicals and assessment of multiple exposure pathways. The major criteria in evaluating these toxicity regimes are related to the mechanisms for incorporating uncertainty and in documenting environmental equity variability across different weighing schemes.

The control or comparison groups used in evaluating environmental equity in different areas also play an important role in determining particular results and in the generalizing of findings. Current research has documented the importance of the dynamic nature of the relationships between TRI impacts and socio-economic characteristics and the need for an historical examination of different areas (Chakraborty and Armstrong, 2001).

## *2.2. Geography of environmental justice*

Most of the social science literature on environmental equity either examines the spatial and, to a lesser extent, the temporal distribution of benefits and burdens (outcome equity) or identifies the causal mechanisms that give rise to these differences in the first place (input and/or process equity) (Cutter et al., 1996). Environmental inequity originates from at least three major sources of dissimilarities – social, procedural, and generational (Bullard, 1994). A large part of the debate surrounding environmental equity appears to be based on (1) the extent of the spatial coincidence between the locations of environmental disamenities and minority or low-income residence (Bowen et al., 1995); and (2) the causal interpretation of how these inequitable relationships developed (Hamilton, 1995).

Several studies on environmental justice have examined the relationship between race and adverse environmental effects by analyzing the relationship between existing pollution sites and current socio-economic, minority, and

population-at-risk characteristics (Chakraborty and Armstrong, 2001). A large proportion of the research tries to map linkages between "race and class" and "environmental exposure and adverse health effects" (Perlin et al., 1999). Several empirical studies have shown that minorities bear a disproportionate share of the burden from the effects of toxic wastes (e.g. Environmental Protection Agency, 1992; Lavelle and Coyle, 1992; Zimmerman, 1993). These studies were conducted for various toxic hazards – TRIs, landfills, and incinerators, National Priority List (NPL) Superfund sites, and waste expansion sites. However, other studies found no statistically significant signs of racial bias to the location of transfer, storage, and disposal facilities (TSDFs) (e.g. Anderton et al., 1997).

Research in environmental equity predominantly concerns the location of various types of environmental hazards in relation to disadvantaged communities. Some authors have examined air quality differences for disadvantaged populations (Freeman, 1972). Others have examined environmental equity in an occupational framework. A great number of efforts have focused on the location, relocation, or expansion of solid or toxic waste sites in relation to different socio-economic populations (Commission for Racial Justice, 1987; Hird, 1993). While some studies have found linkages between race and environmental risks, others find no statistically significant relationships. Others have examined levels of toxic releases and chemical toxicity, and assessed their degree of correlation with respect to disadvantaged populations (Bowen et al., 1995).

The current literature's conception of "spatial" is largely limited to proximity to a site whether it is a TSDF, a TRI site, an NPL Superfund waste clean-up location, or a site on the Comprehensive Environmental Response Compensation and Liability Information System (CERCLIS) location. They do not usually account for spatial phenomena such as the clustering of toxic sites (by Standard Industrial Classification), of distinct at risk populations (age, racial, and/or ethnic cohesion), or by interdependent economic associations (income levels, rents, or property values).

Most researchers present four major hypotheses with regard to the causes of differing environmental quality preferences and decision-making related processes:

- income effects, e.g. market processes leading to lower preferences for environmental quality;
- absence of information by less knowledgeable groups with regard to awareness of health impacts;
- lack of political organization and therefore less ability to influence and be represented in decision-making institutions;
- environmental racism whereby actors with more influence over siting decisions avoid wealthy majority population areas and impose locally unwanted land uses on poor and/or minority groups (Hamilton, 1995; Pollock and Vittas, 1995).

While resolution of this debate is not complete, support for more neutral market-based processes (consistent with theories of industrial location and urban social structure) influencing siting decisions and location of TRI facilities have been articulated (Been, 1994a,b). However, there is some evidence that even when siting is not discriminatory *per se*, location and concentration of facilities can lead to a shifting distribution of more poor and minority populations to these areas. This later notion is consistent with theories and research focusing on locational decisions with regard to transportation and land use patterns within urban areas (Llewellyn, 1981) and migration responses to segregation and forced integration. Thus, the presence of environmental inequities may be tied to market-driven processes in a dynamic manner, where siting decisions are influenced by the nature of the transportation and land use nexus within these urban environments (Mills and Neuhauser, 2000).

### 3. Transportation and environmental justice

The majority of the small but growing research on transportation and environmental justice has taken either a legal and project-specific environmental impact focus or a social cost approach. Kennedy (2000) discusses existing legal cases with regard to transportation planning and environmental justice. As with previous concerns with regard to placement of transportation infrastructure and displacement of low-income/minority groups, these cases have sensitized the transportation planning community with regard to the possible impacts of transportation on exposure to environmental hazards (Freeman, 1972; Federal Highways Administration, 1998, 2000). Under current US federal legislation and laws, metropolitan planning organizations are required to identify and minimize any environment-related impacts of planned transportation facilities and investments.

State agencies and planning organizations also have been urged to incorporate impact assessment methodologies to make transportation more sustainable as well as to address issues of sprawl and access to jobs. Indirect human impacts through land and economic effects of transportation networks and access in specific geographic contexts often occur through developmental factors (Rabin, 1989). This has been recognized in terms of transfer and displacement effects of specific transportation projects on the environment in general, but there is also some recognition of human environmental impacts resulting from these projects. Most of these assessments are consistent with equity planning and efforts to incorporate the direct and indirect social and economic costs and benefits in project analysis in existing transportation policy (Metzger, 1996).

Forkenbrock and Schweitzer (1999) present one of the few systematic efforts to examine and measure the effects of transportation system changes on environmental justice. Their research measures the air quality and noise consequences of US Highway 63 in Waterloo, Iowa, and its impact on poor and minority populations. Using geographic information systems (GIS) and traffic and air pollution models, their methodology holds some promise for the assessment of the relationships between transportation and the location-specific concerns of environmental hazards from TRI facility sitings. However, their prototype application also shows that any environmental inequities in terms of direct and indirect health impacts from different types of release is very localized and requires more detailed analysis at the block and block group levels.

Chen (1997) explores the interaction of social equity, environment, land use, economic development, and transportation in promoting livable communities. He identifies the distributive impacts of transportation investments on communities and their social equity implications. He argues that there is a growing social and environmental cost of mistaken policies in transportation planning. Further, low-income and ethnic minority communities bear the brunt of these costs while enjoying few benefits because of racism, classism, exclusion from the policy-making process, and other forms of discrimination (Pulido et al., 1996).

Wright (1997) chronicles the historical development of New Orleans, and argues that mistaken transportation policies have severely destabilized the city's African-American business corridors and neighborhoods. In her study, Wright illustrates how freeway construction projects contributed to neighborhood destruction in her home city.

To some degree these issues of negative land use characteristics and residential population proximity may be primarily a US urbanization issue, where zoning and land use controls are notoriously weak and unstable. However, as Brown (2003) has reported from the first comprehensive survey for the UK Treasury, over 200 000 homes close to landfill sites have depressed land values averaging GB £5 500 (US \$9 000) per house. This loss of an average 7% of their values ranged regionally to as high as 41% in Scotland. Overall this represented a loss of GB £2.5 billion (US \$4 billion) in the UK due to proximity to landfills. Losses declined with distances.

From empirical urban rent studies we understand that a great deal of both residential and land use distributions are determined in a market context that in turn is highly dependent on access and hence transportation patterns. This means that the central role of transportation infrastructure is a determinative characteristic. In societies with stronger planning and land use zoning traditions the intermediate role of transportation determining land use may be significantly reduced. In the USA, however, that simultaneity of interaction and hence interdependence is palpable.

#### 4. Empirical analysis

In earlier work in this area (Haynes et al., 2001), GIS techniques were used to examine if any spatial patterns can be discerned from relatively raw data using Cuyahoga County (Cleveland, Ohio).

There appeared to be clusters of TRI facilities. Further, socio-economic variables such as the proportion of black population and owner-occupied housing appeared to be spatially concentrated or clustered as well. For example, the proportion of black population is concentrated along a north-east-south-west corridor in Cuyahoga County. While there appeared to be considerable coexistence of TRI facilities and black population, housing vacancy rates, and renter-occupied housing, it was difficult to determine the degree of spatial coincidence, let alone causality, by visually examining spatial associations. Further, TRI locations seem to be clustered to a greater extent in block groups with a higher proportion of renter-occupied units than in predominantly black neighborhoods.

Figure 1 presents a very interesting story. All TRI sites in the metropolitan area appear to be located within 1 mile (1.61 km) of limited-access highways or primary-access roads. These spatial patterns would not be easily discernible if TRI facilities were randomly distributed. This suggests that the location of TRI sites might be motivated to some extent by the ease of access to major transport corridors. Given that income and housing dynamics are also influenced by the location of transport corridors, it is quite possible that we might observe common clustering of TRI facilities and socio-economic variables.

Given that TRI sites are concentrated within 1 mile (1.61 km) of a limited-access highway or a primary road, it would appear that transport access routes are important in TRI location. This is consistent with industrial location theory, which suggests that firms will try to minimize transport costs. The interaction between infrastructure location and industrial pollution is a major concern in the USA, and also has important implications for rapidly developing countries investing in new infrastructure. This is an under-researched area which has the potential to provide critical insights into the location of TRI facilities (Bryant, 1997).

We now focus on the primary planning districts that make up the city of Cleveland. In these districts there are 726 census block groups and 321 TRI sites. These 321 TRI sites are clustered in 79 census block groups.

Figures 2–4 present the clustering patterns of TRI release sites with a slightly increasing distance from the transportation network. Figure 2 displays TRI release sites that are located within a sixteenth of a mile (0.10 km) of a major transportation element. Hollow circles with a black dot at the center represent TRI release sites located within 0.0625 mile (0.10 km) of a major transportation corridor, and solid circles represent the rest of the TRI release sites. The same symbols are used in Figures 3–5 but the distance ranges are different in each case.

Of the 321 TRI release sites, 147 are located within 0.0625 mile (0.10 km) of the transportation buffer in Figure 2. As seen in Figures 2–4, 200 of the sites are within an eighth of a mile (0.20 km) of a major transport link; 263 of the TRI sites are within a quarter of a mile (0.40 km); and 318 of them are within half a mile (0.8 km).

As seen in Figure 5, most of the TRI sites are located within 0.5 mile (0.81 km) of a primary major transportation element. Given that 318 of 321 TRI sites are located within 0.5 mile (0.81 km) of a highway or primary road, the existence of a well-developed transportation network appears to dominate the location decision for siting a TRI facility. Also, previous research on income and housing dynamics

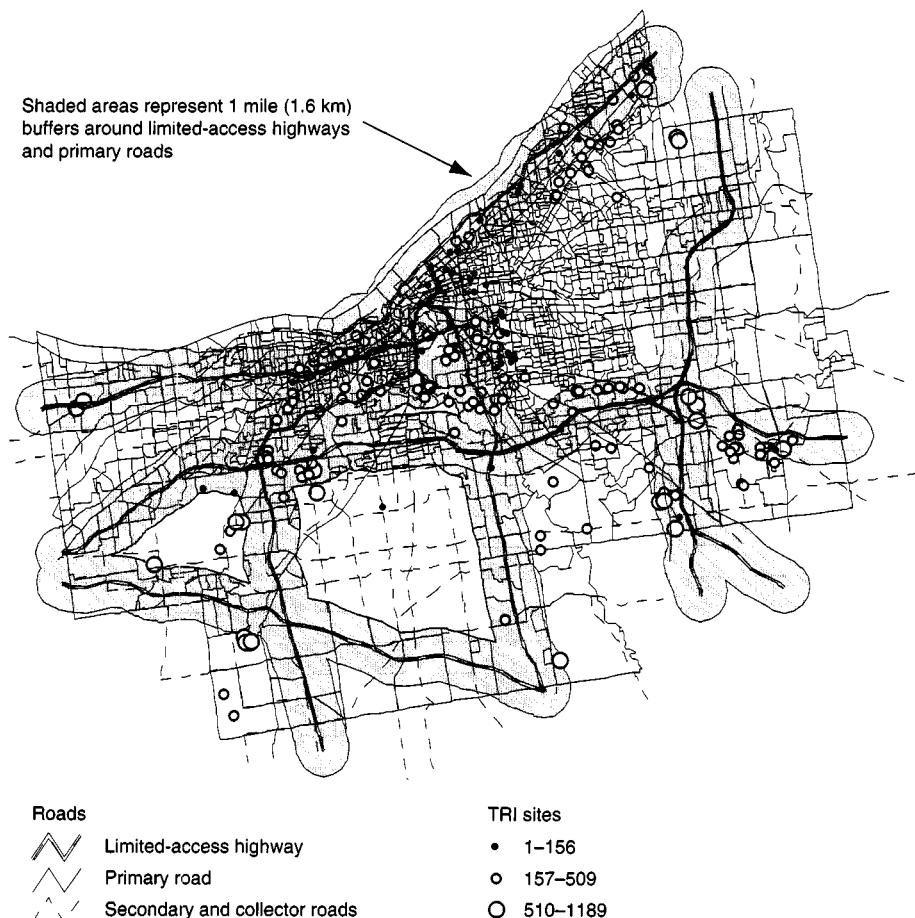


Figure 1. TRI site location and access to transport corridors.

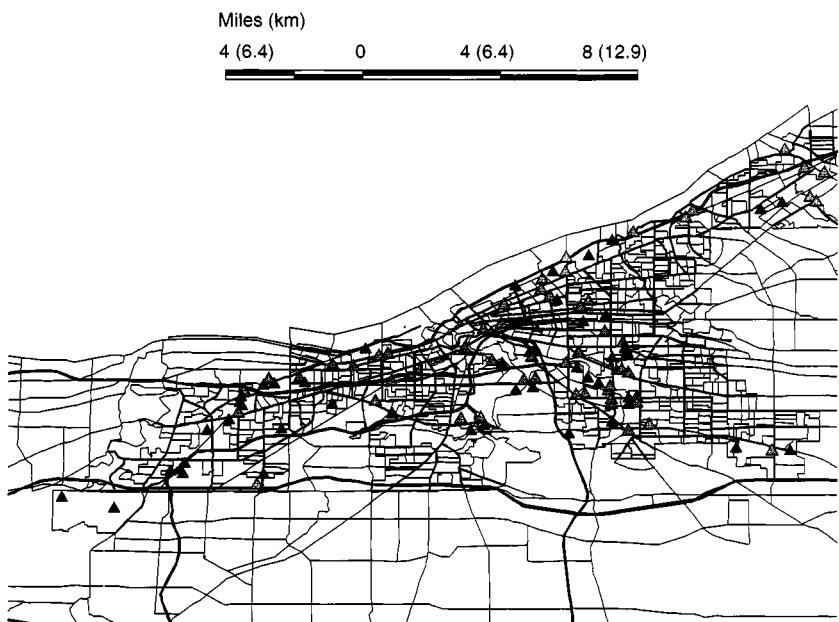


Figure 2. TRI sites within 0.0625 mile (0.10 km) of a major transportation corridor (▲).



Figure 3. TRI sites within 0.125 mile (0.20 km) of a major transportation corridor (▲).



Figure 4. TRI sites within 0.25 mile (0.40 km) of a major transportation corridor (▲).

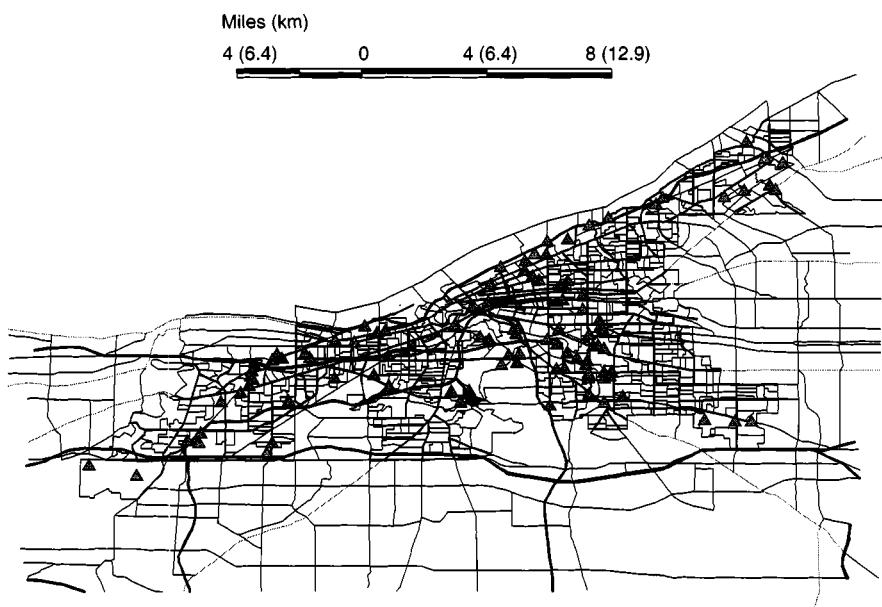


Figure 5. TRI Sites within 0.5 mile (0.81 km) of a major transportation corridor (▲).

Table 1  
Number of TRI sites within specified distances from highways

Distance from transportation corridors (miles (km))	1990		1996	
	Total number of TRI sites	Accumulative percentage	Total number of TRI sites	Accumulative percentage
<0.0625 (0.10)	127	41	147	46
< 0.125 (0.20)	160	52	200	63
<0.375 (0.60)	239	78	263	82
<0.25 (0.40)	298	98	318	99
>0.625 (1.00)	305	100	321	100
Total	305	100	321	100

shows that the location choice of transport influences other socio-economic dynamics (Wright, 1997).

The analysis presented in the previous work in the study area does not find any statistically significant relationship between minority populations and the location of TRI sites in Cuyahoga County, Cleveland, at the census block group level (Haynes et al., 2001). This is consistent with the census tract study of Bowen et al. (1995) at the higher level of spatial aggregation. However, the GIS analysis shows that housing market characteristics and access to transport networks are closely associated with TRI locations. As documented by Haynes and co-workers, most of the TRI sites are located within 0.5 mile (0.81 km) of a highway or primary road.

Table 1 reports the intensity of these findings by noting the number of TRI sites in different distance bands from major transportation corridors. As seen in the table, 98% of the TRI sites were located within 0.25 mile (0.40 km) of a transportation corridor in 1990. This figure increased to 99% in 1996.

Figure 6 shows the percentage distribution of the number of TRI sites across different distance intervals from a major transportation corridor. The curve with solid circles represents the year 1990, and the curve with triangles represents 1996. In order to analyze the general pattern of the distribution of TRI sites, an exponential trend line is estimated for each year and presented in Figure 6 (the broken line is the trend line for 1996 and the solid line is the trend line for 1990). An exponential trend line is a curved line that is most useful when data values rise or fall at increasingly higher rates – which is the case in the cumulative distribution of TRI sites. The simple exponential trend line is calculated by using least square estimators fitted by  $y = ce^{bx}$ , where  $c$  and  $b$  are constants.

As seen in Figure 6, the trend lines are located to the right of the actual distribution curves for both years, which suggests that there is a difference between the fitted and the actual distribution patterns of TRI sites. The actual

distribution of TRI sites shows a strong pattern of clustering around major transportation bands, with higher percentages than the trend line predicts. Also, the upper ends of the predicted and actual curves for 1996 are shifting upward to the left in comparison with those for 1990, which suggests that more of the TRI sites are located closer to the transportation corridors in 1996 than in 1990.

Figure 7 shows the distribution of the volumes of toxic releases and transfers. The same trend as in Figure 6 can be observed with regard to these observations. Table 2 documents the cumulative proportion of volumes of toxic materials released and transferred within the same distance intervals from a major transportation corridor as in Table 1. In 1990, 17% of the total volume of toxic releases and transfers occurred within 0.0625 mile (0.10 km) of a major transportation corridor. This increased to 19% in 1996. In fact, the cumulative percentages increased for all the categories except for the open upper limit, i.e. 0.5 mile (0.81 km) and over, as shown in Figure 7.

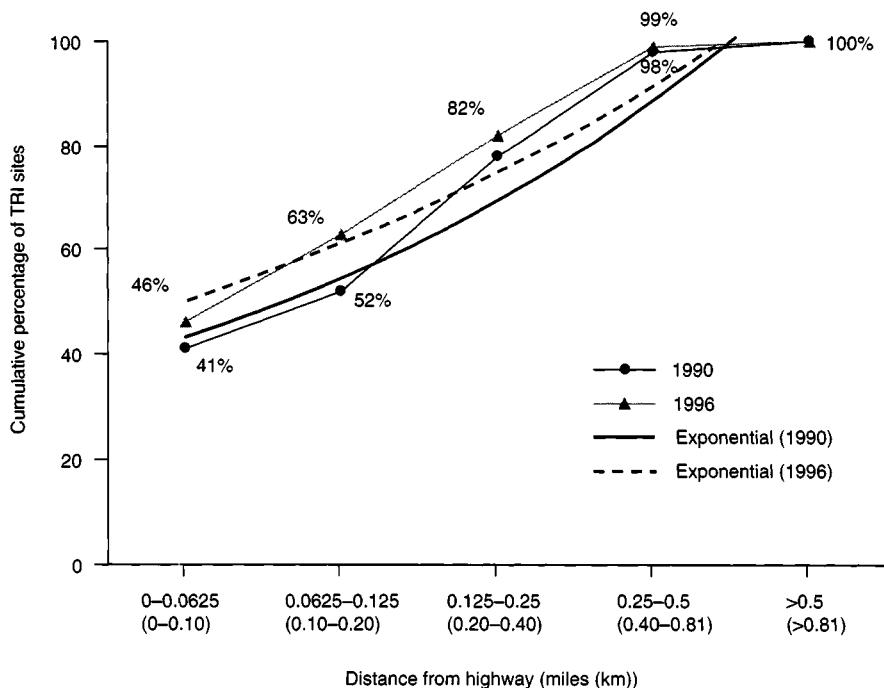


Figure 6. Distribution of TRI sites.

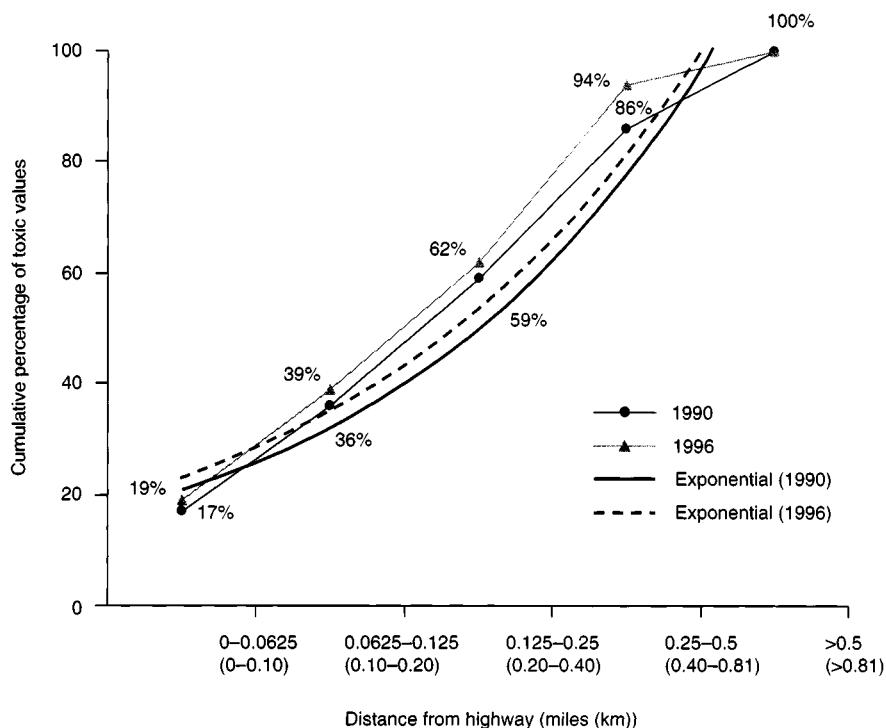


Figure 7. Distribution of toxic release and transfer volumes.

Table 2  
Volumes of total release and transfer of toxic materials within specified distances from highways

Distance from transportation corridors (miles (km))	Cumulative percentage of toxic materials released and transferred	
	1990	1996
<0.0625 (0.10)	17	19
< 0.125 (0.20)	36	39
<0.375 (0.60)	59	62
<0.25 (0.40)	86	94
>0.5 (0.81)	100	100

Table 3  
Average distance of TRI sites from transportation corridors in 1996

Distance interval (miles (km))	Number of TRI sites ( $F_i$ )	Mid-point ( $M_i$ ) (miles (km))	$F_i M_i$
0–0.0625 (0–0.10)	147	0.0313 (0.0504)	4.5938 (7.4088)
0.0625–0.125 (0.10–0.20)	53	0.0938 (0.1510)	4.9688 (8.0030)
0.125–0.25 (0.20–0.40)	63	0.1875 (0.3018)	11.8125 (19.0134)
0.25–0.5 (0.40–0.81)	55	0.3125 (0.5029)	17.1875 (27.6595)
0.5–1.0 (0.81–1.61)	3	0.7500 (1.2070)	2.2500 (3.6210)
Total	321	0.1270 (0.2044)	40.8125 (65.6124)

For environmental justice to occur as transportation system investments are made, their probable consequences must be estimated and the results presented comprehensibly to affected populations. This is difficult because most infrastructure investment studies focus on immediate local impacts rather than longer-term restructuring consequences of such investments. A growing awareness of the disproportionate exposure of poor and minority groups to environmental disamenities has led community activists to argue that toxic release and transfer sites are chosen based on the local population characteristics of these locations. However, environmental justice is a multifaceted public policy problem. As argued in this chapter, transportation investments and the existence of TRI sites and toxic release levels are inexorably connected. The example of Cleveland shows that the number of TRI sites and the intensity of releases are highly concentrated in proximity to the transportation network, which suggests that it is the transportation network that significantly affects the decisions of toxic wastes siting. Given the current spatial distribution of TRI sites, the probability that new TRI sites would be located more than 0.5 mile (0.81 km) away from a primary transportation network is very slim (Table 3).

In 1996 the average distance from a TRI site to a major transportation corridor was only  $40.8125/321 = 0.127$  mile (0.204 km), with a standard deviation of 0.0146 mile (0.0235 km). Using the normal distribution, the probability can be determined by computing the  $Z$  value:

$$\begin{aligned}
 Z &= (x - \bar{x})/\sigma \\
 &= (0.5 - 0.127)/0.0146 \\
 &= 25.55.
 \end{aligned}$$

A  $Z$  value of 25.55 corresponds to a probability of less than 0.001. This means that,

given the distribution of TRI sites along the transportation network, we are more than 99% sure that a new TRI site would locate within 0.5 mile (0.81 km) of a transportation corridor. Further, the siting results are reinforced by using pollution load information, indicated by changes in volumes of release and transfer at these sites. This indicates that site use levels also increase with proximity to major road links.

## 5. Conclusion

The connection between transportation and land use is a fundamental concept in transportation planning. Social and economic research related to transportation is a broad subject, because of the many ways in which transportation policies and actions can produce known social and economic impacts on populations. At the same time, transportation investment decisions can produce unforeseen and unintended impacts on the population as well.

The concept of environmental justice and the design of transportation networks are intrinsically related even when a toxic or noxious facility is not intentionally sited in or near poor or minority residential areas. This study presents evidence in which location of TRI sites is affected by transportation consideration. Transportation investments change the access to urban space, and as a consequence economic decisions related to time and route cost minimization are then reflected in downstream location and utilization decisions. Such decisions include the location and utilization of public and private waste facility infrastructures.

The incidence of pollutants needs to be examined carefully with distance decay models of polluting agents and development of toxicity indices (e.g. Harner et al., 2002) to appreciate the risks associated with specific pollutants. This process should make it possible to get a better picture of the risks associated with the release of specific industrial pollutants as well as the populations that are vulnerable to these releases.

The temporal sequence of TRI sites and other socio-economic characteristics such as minority groups, low-income populations, and renter-occupied housing requires further studies. There have been some studies of this kind, suggesting that minorities and low-income populations followed the establishment of waste and hazardous facilities (Mitchell et al., 1999). However, it is necessary to have more studies that are designed for various types of areas and regions representing different evolutionary routes of urban land use processes. A link between the study of environmental justice and the economics of residential location, particularly as it relates to transportation, should provide some insight in this regard. A significant contribution in this respect is the Chicago study, which

suggests that historical transportation investments, even canals, have very long-term redistributive effects (Baden and Coursey, 2002).

Further, the approach to environmental risk assessment *per se* has important implications for the study of environmental justice. Very recently, a shift in environmental assessment from risk to exposure assessment has been proposed as an approach to address the challenges from environmental justice (Corburn, 2002). As is stated in a report by the National Academy of Public Administration (2001), environmental justice is now an integral part of the US Environmental Protection Agency policy, and as such it is a central federal policy concern that needs to be more effectively addressed. Thus, a change or modification of the existing environmental assessment process should be considered on the basis of further analysis.

In summary, the environmental justice literature suggests that the poor and minorities are exposed disproportionately to environmental disamenities in both rural and urban environments. The transportation literature has noted a historical pattern of urban placement of its facilities through minority and low-income neighborhoods, as a result of social policy in the 1970s related to "slum clearance," as a result of market mechanisms utilizing lower-valued lands, and as a result of weak and/or poorly organized community resistance. Hence, the negative environmental externalities from transportation system corridors such as traffic safety problems and noise and air pollution historically have become coterminous with the poor and minority communities through which they traverse. The findings reported in this chapter suggest an additional subtle and continuing link between transportation and negative environmental externalities. After decades of massive transport investment in the USA the patterns of urban industrial location have been reorganized, resulting in increased mobility and more efficient movement of goods and people on the one hand and a redistribution of environmental disamenities on the other. Positive transport externalities include a widened access to suppliers, markets, and labor, and a corresponding change in the location and utilization of public and private infrastructure investments that are intended to manage negative industrial outputs. The infrastructure for managing these negative industrial outputs or waste include landfills, incinerators, TSDFs that show up on information collected as part of the TRI, NPL Superfund sites, and the CERCLIS. The link and interdependence between infrastructure investments is central to our understanding and appreciation of the continuing nexus between transportation and environmental justice. The example findings presented in this chapter represent a step toward understanding the benefits, costs, and long-term consequences of investments in transportation infrastructure. These findings, along with research needs in other aspects of environmental justice, present a research agenda that merits significant effort and attention in the coming years.

## Acknowledgments

The author gratefully acknowledges the financial support of NSF/EPA grant SES-9976483 “Social vulnerability analysis: spatial and hazard screening of toxic chemical releases” and NSF grant ECS-0085981 “Road transportation as a complex adaptive system.” Assistance by Sandra Henry and Qingshu Xie is appreciated.

## References

- Anderton, D.L., J.M. Oakes and K.L. Egan (1997) “Environmental equity in Superfund: demographics of the discovery and prioritization of abandoned toxic sites,” *Evaluation Review*, 21:3–26.
- Been, V. (1994a) “Locally undesirable land uses in minority neighborhoods: disproportionate siting or market dynamics,” *Yale Law Journal*, 103:1383–1422.
- Been, V. (1994b) “Market dynamics and the siting of LULUs: questions to raise in the classroom about existing research,” *West Virginia Law Review*, 96:1069–1078.
- Bowen, W.M. (2001) *Environmental justice through research-based decision-making*. New York: Garland.
- Bowen, W.M. and K.E. Haynes (2000a) “Environmental injustice: is race or income a better predictor?” *Social Science Quarterly*, 81:885–888.
- Bowen, W.M. and K.E. Haynes (2000b) “The debate on environmental injustice,” *Social Science Quarterly*, 81:892–894.
- Bowen, W.M., M.J. Salling, K.E. Haynes and E.J. Cyran (1995) “Toward environmental justice: spatial equity in Ohio and Cleveland,” *Annals of the Association of American Geographers*, 85: 641–663.
- Braden, B.M. and D.L. Coursey (2002) “The location of waste sites within the city of Chicago: a demographic, social, and economic analysis,” *Resource and Energy Economics*, 24:53–93.
- Brown, P. (2003) “How a landfill site can hurt the value of your home,” *The Guardian*, Feb. 27:10.
- Bryant, B. (1995) “Pollution prevention and participatory research as a methodology for environmental justice,” *Virginia Environmental Law Journal*, 14:589–613.
- Bryant, B. (1997) “Environmental justice, consumption, and hazardous waste within people of color communities in the US and developing countries,” *International Journal of Contemporary Sociology*, 34:159–171.
- Bullard, R.D. (1994) *Dumping in Dixie: race, class, and environmental quality*, 2nd edn. Boulder: Westview Press.
- Chakraborty, J. and M.P. Armstrong (2001) “Assessing the impact of airborne toxic releases on populations with special needs,” *Professional Geographer*, 53:119–131.
- Chen, D. (1997) “Linking social equity with livable communities,” in: R.D. Bullard and G.S. Johnson, eds, *Just transportation: dismantling race and class barriers to mobility*. New York: New Society Publishers.
- Collins, R.W. (1992) “Environmental equity: a law and planning approach to environmental racism,” *Virginia Environmental Law Journal*, 11:495.
- Commission for Racial Justice (1987) *Toxic wastes and race: a national report on the racial and socioeconomic characteristics of communities with hazardous waste sites*. New York: United Church of Christ.
- Corburn, J. (2002) “Environmental justice, local knowledge, and risk: the discourse of a community-based cumulative exposure assessment,” *Environmental Management*, 29:451–466.
- Cutter, S.L. (1995) “Race, class and environmental justice,” *Progress in Human Geography*, 19: 111–122.

- Cutter, S.L., D. Holm and L. Clark (1996) "The role of geographic scale in Monitoring environmental justice," *Risk Analysis*, 16:517–526.
- Environmental Protection Agency (1992) *Environmental equity – reducing risk for all communities*. Washington, DC: EPA.
- Federal Highway Administration (1998) *Community impact mitigation: case studies*, FHWA-PD-98-024 HEP30/5-98. Washington, DC: US Department of Transportation.
- Federal Highway Administration (2000) *Environmental justice: an overview of transportation and environmental justice*, FHWA-EP-00-013. Washington, DC: US Department of Transportation.
- Forkenbrock, D.J. and L.A Schweitzer (1999) "Environmental justice in transportation planning," *Journal of the American Planning Association*, 65:96–111.
- Freeman, M. (1972) "The distribution of environmental quality," in: A.V. Kneese and B.T. Bower, eds, *Environmental quality analysis: theory and methods in the social sciences*. Baltimore: Johns Hopkins University Press.
- Greenberg, M. (1993) "Proving environmental inequity in siting locally unwanted land uses," *Risk: Issues in Health and Safety*, 4:235–252.
- Hamilton, J.T. (1995) "Testing for environmental racism: prejudice, profits, political power?" *Journal of Public Policy Analysis and Management*, 14:107–132.
- Harner, J., K. Warner, J., Pierce and T. Huber (2002) "Urban environmental justice indices," *Professional Geographer*, 54:318–331.
- Haynes, K.E., S.V. Lall and M.P. Trice (2001) "Spatial issues in environmental equity," *International Journal of Environmental Technology and Management*, 1:17–31.
- Hird, J. (1993) "Environmental policy and equity: the case of Superfund," *Journal of Public Policy Analysis and Management*, 12:323–343.
- Kennedy, L.G. (2000) "Environmental justice and where it should be addressed in the 21st century concerning the transportation industry: historical perspective and summary," in: *Conference proceedings 20: refocusing transportation planning for the 21st century*. Washington, DC: Transportation Research Board.
- Kunreuther, H. and D. Easterling (1996) "The role of compensation in siting hazardous facilities," *Journal of Public Policy Analysis and Management*, 15:601–622.
- Lavelle, M. and M. Coyle (1992) "Unequal protection: the racial divide in environmental law," *National Law Journal*, Sept, 21:S1.
- Llewellyn, L. (1981) "The social costs of urban transportation," in: I. Altman, J.F. Wohlwill and P.B. Everett, eds, *Transportation and behavior*. New York: Plenum Press.
- Metzger, J. (1996) "The theory and practice of equity planning: an annotated bibliography," *Journal of Planning Literature*, 11:112–126.
- Mills, G.S. and K.S. Neuhauser (2000) "Quantitative methods for environmental justice assessment of transportation," *Risk Analysis*, 20:377–395.
- Mitchell, J.T., D.S.K. Thomas and S.L. Cutter (1999) "Dumping in Dixie revisited: the evolution of environmental injustices in South Carolina," *Social Science Quarterly*, 80:229–243.
- National Academy of Public Administration (2001) *Environmental justice in EPA permitting: reducing pollution in high-risk communities is integral to the agency's mission*. Washington, DC: NAPA.
- Perlin, S.A., K. Sexton and D.S. Wong (1999) "An examination of race and poverty for populations living near industrial sources of air pollution," *Journal of Exposure Analysis and Environmental Epidemiology*, 9:29–48.
- Pollack, P.H. and E.M. Vittas (1995) "Who bears the burden of environmental pollution? Race equity in Florida," *Social Science Quarterly*, 76:294–309.
- Pulido, L., S. Sidawi and R.O. Vos (1996) "An archaeology of environmental racism in Los Angeles," *Urban Geography*, 17:419–439.
- Rabin, Y. (1989) "Expulsive zoning: the inequitable legacy of Euclid," in: C. Harr and J. Kayden, eds, *Zoning and the American dream: promises still to keep*. Washington, DC: American Planning Association Press.
- Stutz, F. (1986) "Environmental impact," in: S. Hanson, ed., *The geography of urban transportation*. New York: The Guilford Press.
- Taylor, D.E. (2000) "The rise of the environmental justice paradigm: injustice framing and the social construction of environmental discourses," *American Behavioral Scientist*, 43:508–580.

- Wright, B.H. (1997) "New Orleans neighborhoods under siege," in: R.D. Bullard and G.S. Johnson, eds, *Just transportation: dismantling race and class barriers to mobility*. New York: New Society Publishers.
- Zimmerman, R. (1993) "Social equity and environmental risk," *Risk Analysis*, 13:649–666.

***Part 2***

**TRANSPORT AND SPATIAL FORM**

## TRANSPORT IN THE URBAN CORE

EVELYN BLUMENBERG and RANDALL CRANE

*University of California at Los Angeles*

### 1. Introduction

On the one hand, the urban core – or central city, or “downtown” – is not a particularly well-defined part of the metropolitan area. Administrative, political, or bureaucratic boundaries rarely draw the lines in a manner that serves our purposes as transportation analysts. The center may be older, denser, and/or poorer than the surroundings, and nearer the compass center, but these distinctions are neither hard nor fast. On the other hand, whatever and wherever it is, the core may well be more like the city at large than not, such that the data, analyses, and conclusions pertaining to the city as a whole apply to core areas too. Put another way, a given center may have less in common with other centers than with the rest of the region it is contained by. It would be interesting to explore this hypothesis by comparing and contrasting the transportation environments of different central cities, and the urban regions they represent, and then to track those trends over recent history. How do these relationships and patterns vary, are they converging across space or time, and with which local features are various outcomes associated? That approach would leave us with a better sense of the different ways that the core fits into the metropolitan travel story, and why.

Limited space does not permit this analysis here, so we instead take the main conventional distinctions as given and explore the implications for transportation problems and policies. That the core does tend to be denser, poorer, and older suggests the key themes of this chapter: density, poverty, and decentralization.

The city center is often the original city, around which the metropolitan area has grown and decentralized. This is clearer in Europe than in much of the USA, or in newer “edge” cities anywhere, but only as a matter of degree. Except where redevelopment has been extensive, this may mean that core facilities are older and in need of maintenance or replacement, or designed to standards less likely to accommodate automobiles – or at least to accommodate them in the numbers they likely must today. Together with high employment and residential densities, this

can lead to higher traffic congestion in central cities than elsewhere. The urban core also tends to be poorer, which has implications for the resources available for transport as well as the role of transport in generating income. And finally, its role as the economic center is evolving.

For these and related reasons, people travel differently in the urban core than in the suburbs: they take longer to commute to work, are more likely to walk and use bus and rail, and drive less often and less far in the USA (US Federal Highway Administration, 2001). Yet the decentralization of employment and population, generally, and the rise of secondary employment centers, in particular, has substantially weakened the center's traditional pull in recent decades. The exodus has not been uniform, with different groups and industries emigrating (and immigrating) at different rates, so that many observers are concerned that the mobility and access of the core may be worsening with decentralization.

That said, the city center also has several traits associated with favorable transport conditions. Density implies proximity. Land uses are less segregated, again suggesting proximity. Proximity in turn implies higher accessibility through both shorter trips and more potential destinations. Pedestrian and transit modes shares tend to be higher, and the hub role played by many metropolitan areas facilitates rail and express bus systems (US Federal Highway Administration, 2001). In many respects, the urban core resembles the highest form of "new urbanist" and transit-oriented designs aimed at reducing motorized travel in general and car use in particular.

While such trends are largely transparent and can be identified in substantial detail with the data available these days, they are also complex to assess. Transport scholars face many unanswered questions about the core. Some are positive, such as the precise manner in which specific features of the built environment, including employment access and land use patterns, influence travel patterns and mode choice. Others are normative: what is the best mix of transport policies, including facility design, means of finance, and land use planning?

This chapter will not answer open questions. Rather, its purpose is to set the stage for useful future study by highlighting what we do know on several key issues. It proceeds by discussing key challenges, promises, and policy options in turn.

## **2. Density**

Congestion is increasing across all urban areas across the world, but through the centuries has been most closely associated with the city center. For example, passenger-kilometers of travel increased over 85% on freeways and major streets, and approximately 25% on transit systems in 75 urban areas studied in the USA from 1982 to 2000. Also during this time, the percentage of lane-kilometers of congested roadway greatly expanded from 34% to 58%, and some estimate

congestion costs upward of US \$68 billion in wasted fuel and lost productivity in the USA alone (Schrank and Lomax, 2002). Although it is difficult to determine from the data how much congestion increased within the urban core in particular, what is clear is that almost all urban residents spend some time stuck in traffic either in their own vehicles or on city buses.

As a measure of how traffic activity matches up with capacity at any point in time, congestion is by most accounts a mixed blessing. For example, high levels of traffic tend to indicate a vigorous economy. Not surprisingly, data for the year 2000 show congestion levels to be significantly higher in vibrant cities, such as Los Angeles, San Francisco, Washington, DC, and San Jose in the USA, relative to more economically depressed metropolitan areas such as Pittsburgh or Cleveland. No congestion can mean no economic activity. It is also the case that congestion is to some extent self-regulating, in that people choose other modes or different times of day or to drive less under congested conditions.

However, congestion is also a classic externality in the sense that individual travelers do not account for how they slow everyone else down. That is, they travel “too much” when oblivious to the social costs of their actions, even while well aware that other drivers slow them down. Hence, demand can be high in part due to economic conditions and in part due to an unregulated externality.

The “downtown” part of this story is twofold: travel demand per square kilometer is greater in higher densities and economic centers; and, on the supply side, streets and road capacity may be more scarce than in the suburbs, relative to demand. US data from the *1995 and 2001 National Personal Transportation Surveys* (US Federal Highway Administration, 1995, 2001) show that density is correlated with transit use; the greater the population density, the higher the percentage of residents who use public transit (Pucher and Renne, 2003).

Interpreting these data is complex since they also reflect the high concentrations of captive riders – largely low-income adults in households without personal vehicles – residing in city centers. However, this relationship persists even when controlling for household income; for example, among households with incomes over US \$75 000 the percentage of commuters reliant on public transit is still very high (21.5%), more than three times as high as for high-income suburban commuters (6%).

What to do? The problem is in part demand, much of which simply reflects a healthy economy. We do not generally want to depress economic conditions simply to ease traffic. Demand is also excessively high when travelers do not appreciate the burdens they place on fellow drivers. The latter can most simply be addressed by way of additional charges during congested times at congested places, as is now done in, for example, central Hong Kong and London.

However, another part of the problem is capacity. The US Federal Highway Administration (2001) reports that between 1980 and 1999, the route-kilometers of highways increased 1.5% while vehicle-kilometers of travel increased 76%. The

mismatch between highway construction and vehicle-kilometers of travel may suggest that central-city residents especially would benefit from either road pricing or capacity expansion – both of which are politically and economically challenging. Travel would also be enhanced by improving the quality of inner-city roads and highways, many of which are old, deteriorating, and hazardous. Although capacity improvements may only temporarily ease congestion as other drivers take advantage of less congested roadways, overall they would increase mobility and, therefore, accommodate greater economic and social activity.

### 3. Poverty

The geography of poverty is a fundamental aspect of the city center, including its transport. The urban core is significantly poorer than the rest of the metropolitan area. In the USA, for example, the spatial concentration of poverty increased steadily through the last three decades of the last century, with over 25% of the core's population made up of the nation's poorest 20% by 1998 (Table 1).

Table 1  
The share of USA central-city population, by income class and year

Year	All metropolitan statistical areas/ primary metropolitan statistical areas (%)	Central cities (%)	Suburbs (%)
<b>Low income (national lowest 20%)</b>			
1969	18.3	21.9	14.8
1979	18.5	23.7	14.5
1989	18.1	24.0	14.1
1998	19.0	25.5	14.9
<b>Middle income (national middle 60%)</b>			
1969	59.4	59.8	59.1
1979	59.4	59.0	59.8
1989	59.4	58.8	59.8
1998	58.8	57.9	59.3
<b>High income (national top 20%)</b>			
1969	22.3	18.3	26.2
1979	22.1	17.3	25.7
1989	22.5	17.2	26.1
1998	22.3	16.6	25.8

Source: US Department of Housing and Urban Development (2000).

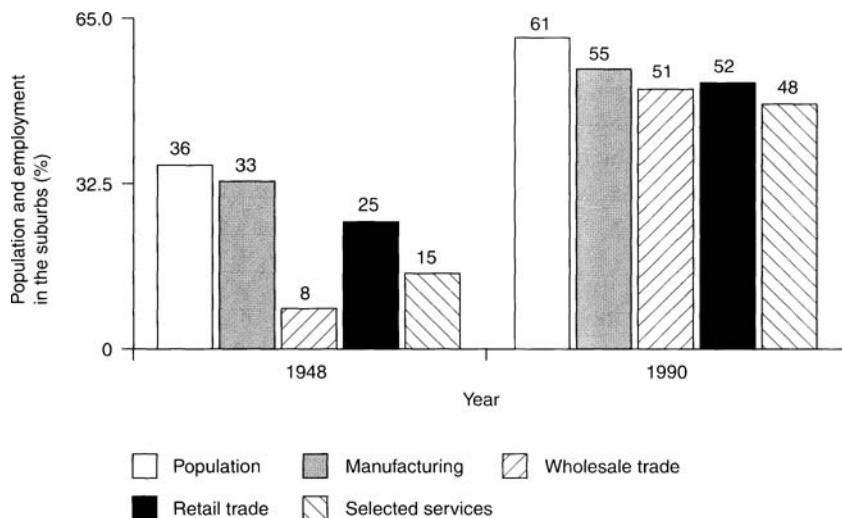


Figure 1. Population and employment decentralization in the USA. (Source: US Census Bureau, 2000.)

Transport/poverty issues take several forms: transportation costs themselves, access to jobs and other destinations, access by modes, and the transportation policies aimed at each. Job access for inner-city residents is compounded by job decentralization, discussed in the next section. Central city job access for suburban residents is clearly a closely related issue.

#### 4. Decentralization

One of the plainest facts of metropolitan areas is the continuing decentralization of both people and their jobs (US Department of Housing and Urban Development, 2000). In 1948 some 64% of the US population lived in the urban core, as defined by the US Census Bureau. By 1990, this share had fallen to less than 40% (Figure 1).

What do these trends portend for the core? We review this issue from two different perspectives: the spatial mismatch literature, which focuses on the impacts and determinants of lengthening commutes for urban core minority residents; and the commute length literature, which asks if employment decentralization is associated with longer commutes after all.

#### *4.1. Spatial mismatch*

People and their jobs are suburbanizing almost everywhere. Or, sometimes, only their jobs. Many low-income inner-city residents increasingly find themselves miles away from employment opportunities. Termed the “spatial mismatch hypothesis,” the spatial separation of residents and housing has been linked to high unemployment rates and growing inner-city poverty (Kain, 1968).

As is often the case, the causal process is not altogether clear. Kain’s main interest was in exploring whether minority households were constrained from following job growth to the suburbs by housing discrimination. That is, whether they stayed in the city center as a matter of choice or not. A voluminous literature followed (Kain, 1992). For our purposes, what is relevant is that the distance between inner-city residences and places of employment is frequently amplified by a lack of adequate transportation. Low-income residents living in the urban core are disproportionately reliant on public transit, which has traditionally done a poor job of moving residents from dense urban areas to dispersed suburban locations. Twenty-six percent of low-income adults in the USA live in households without personal vehicles (Murakami and Young, 1997). Unless they are able to borrow cars or ride with other drivers, these adults are likely to be dependent on public transit. And while transit works best in the inner city, where there are dense clusters of jobs and residents, it has difficulty serving the reverse commute – from central city to suburb – where employment is less frequently located adjacent to transit stops (Cervero et al., 2002).

The spatial mismatch hypothesis is an important concept, and while it has been put to the test by a host of skeptical scholars – some of whom have been able to challenge aspects of the theory – ample empirical support for the concept remains (Holzer, 1991; Kain, 1992; Ihlanfeldt and Sjoquist, 1998; Preston and McLafferty, 1999). However, for transportation policy, it is important to acknowledge that the rapid migration of jobs toward the suburbs does not suggest that there are no employment opportunities in central cities. Some metropolitan areas – for example in the north-east and mid-west USA – have experienced a dramatic hollowing out of the urban core.

However, this experience is far from universal. The decline in central-city Detroit or Cleveland looks vastly different from that of Las Vegas, New York, Phoenix, or even Los Angeles.

With respect to the inner-city poor, residents who find employment tend to find it in the central city, since the average commute distance among low-income workers to central city jobs is relatively shorter (Khattak and Quercia, 2000). And in some metropolitan areas there are more entry-level job opportunities in the central city than in the suburbs, particularly if job turnover rather than job growth is used as the measure of job availability (Shen, 1998, 2001). However, in these cases, the problem for unemployed workers may lie in the relative competition for

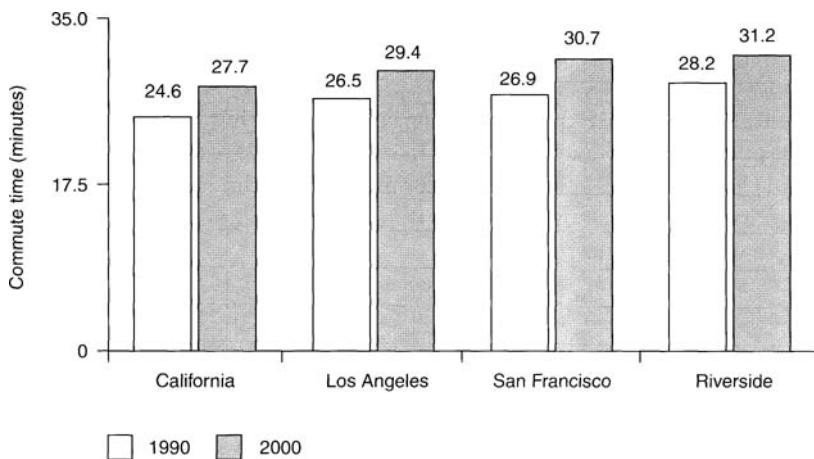


Figure 2. Commute times in California. (Source: US Census Bureau, 2000.)

available jobs. In densely populated inner cities, many potential workers vie for each opening, reducing the chances of employment.

#### *4.2. The journey to work*

Employment is decentralized, but at different rates by sector. The most centralized in 1948 were wholesale trades, at 92%. One of the least was manufacturing, of which 67% was located in the central city. By 1990, central wholesale trade employment was less than half of the total urban share, and the central employment share for manufacturing was less than 45%.

A popular view is that employment decentralization has lengthened the average journey to work, and that, together with congestion, has increased average commute times. Indeed, the observed trend is that in many metropolitan areas, commute times have risen – as shown by the US Census Bureau data for California in Figure 2. However, since workers are also leaving the core, we do not know whether the contributing role of employment suburbanization has been to exacerbate or mitigate this trend.

Still, 39% of all commutes in major US metropolitan areas are within the central city, and another 12% are from the suburbs into the central city. The central city is the destination for 51% of all commuters, significantly higher than the 36% who travel within the suburbs (Pisarski, 1997).

In an important study 15 years ago, Gordon et al. (1989) used county-level data for US metropolitan areas to investigate this question. They found that commutes

in spatially larger cities took more time, while shorter commutes were associated with a higher proportion of industrial employment. Both higher overall residential density and the share of employment in the central city were highly associated with commutes of longer duration. They concluded that both residential and employment dispersion leads to shorter commutes. More recent research by Crane and Chatman (2003), using a national panel for the USA spanning 1985 to 1997, supports this in part, but indicates that average commutes depend on which industries have suburbanized.

Thus, on the one hand transportation access in the core is a problem both for (1) workers living in the core whose jobs are suburbanizing, and who cannot follow for one reason or another, and for (2) workers residing in the suburbs who work in the center, but who face congested traffic conditions. Employment suburbanization offers relief for the second group, but it would seem at the expense of the first.

That said, existing roads and highways in the urban core currently provide automobile users with relatively good access to employment and services within a reasonable commute time in the urban USA. For example, in Cleveland, residents living in the urban core could drive to over 40% of the employment opportunities in the area within a 30 minute commute (Leete and Bania, 1999). Similarly, in Los Angeles, a 30 minute commute by automobile allows inner-city residents access to most of the employment in the area (Blumenberg and Ong, 2001).

In contrast, both of the aforementioned studies find that the transit-dependent suffer heavily compared with those reliant on personal vehicles. Even in areas with high concentrations of employment, residents with access to automobiles can reach as many as five times more employment opportunities within a 30 min commute than those who are transit-dependent (Blumenberg and Ong, 2001).

The purpose of the transportation infrastructure of a city is to increase the mobility of residents, to improve access to an array of needed destinations – the workplace, social services, leisure activities, and so on. Given current land use patterns and, in particular, the dispersion of employment opportunities, the automobile will remain an important mode of transportation even within the central city. However, fixed-route public transit could play a greater role if it concentrated on doing what it does best – serving densely developed corridors. Most public transportation systems have been designed for middle-class suburban commuters heading inbound to the central business district. Efforts to entice higher-income commuters on to public transit have led to an emphasis on expensive rail programs at the expense of buses, used primarily by low-income and minority residents (Garrett and Taylor, 1998–1999). These priorities have resulted in a mismatch between the transit demands of central-city residents and the capacity of inner-city bus systems to accommodate this demand, helping to explain many of the complaints frequently lodged against using buses: lengthy travel times, long waits at bus stops, and unreliability.

Similarly, low-income, central-city job seekers would benefit from transit services that provide increased access to distant neighborhoods where job competition is less severe. Currently, most of these commutes – even among low-income workers – are made in private vehicles (Cervero et al., 2002). More recently, agencies have been experimenting with demand-responsive programs such as vanpools and shuttles that provide low-income, inner-city residents with door-to-door transit service to suburban employment opportunities (Cervero et al., 2002).

## 5. Summary

How are transport problems and solutions different in the urban core to those in the suburbs? We have attempted to answer this question by emphasizing density and access issues that underlay a number of prominent scholarly debates, from the spatial mismatch of jobs and housing, to the role of density in reducing automobile dependence, to the feasibility of public transit for providing employment access for low-income workers. Central city residents and workers may travel differently, on average, to their suburban counterparts, but many indicators suggest that these differences are both nuanced and vanishing. In a world of rapidly decentralizing employment, the car is increasingly king.

Still, life in the urban core is less dependent on the automobile than elsewhere, and rising congestion will continue to encourage the use of other modes. In particular, city centers have virtually all the key traits that new urbanist advocates aspire to, including high residential and commercial densities, mixed land uses, transit access, pedestrian access, and so on.

In the end, the urban core is much like the rest of the city, only more so: the transport challenges are more acute; and the opportunities that much more valued.

## References

- Blumenberg, E. and P. Ong (2001) "Cars, buses, and jobs: welfare recipients and employment access in Los Angeles," *Journal of the Transportation Research Board*, 1756:22–31.
- Cervero, R., Y. Tsai, M. Wachs, E. Deakin, J. Dibb, A. Kluter, C. Nuworsoo, I. Petrova and M. Pohan (2002) *Reverse commuting and job access in California*. Los Angeles: California Department of Transportation.
- Crane, R. and D. Chatman (2003) "Job sprawl and the journey to work in the U.S.A.: 1985–1997," in: C. Bae and H. Richardson, eds, *Sprawl in Western Europe and the United States*. London: Ashgate.
- Garrett, M. and B. Taylor (1998–1999) "Reconsidering social equity in public transit," *Berkeley Planning Journal*, 13:6–27.
- Gordon, P., A. Kumar and H. Richardson (1989) "The influence of metropolitan spatial structure on commuting time," *Journal of Urban Economics*, 26:138–151.

- Holzer, H.J. (1991) "The spatial mismatch hypothesis: what has the evidence shown?" *Urban Studies*, 28:105–122.
- Ihlanfeldt, K.R. and D.L. Sjoquist (1998) "The spatial mismatch hypothesis: a review of recent studies and their implications for welfare reform," *Housing Policy Debate*, 9:849–892.
- Kain, J.F. (1968) "Residential segregation, Negro employment, and metropolitan decentralization," *Quarterly Journal of Economics*, 82:175–197.
- Kain, J.F. (1992) "The spatial mismatch hypothesis: three decades later," *Housing Policy Debate*, 3:371–460.
- Khattak, A.J. and R. Quercia (2000) "Are travel times and distances to work greater for residents of poor urban neighborhoods?" *Transportation Research Record*, 1718:73–82.
- Leete, L. and N. Bania (1999) "The impact of welfare reform on local labour markets," *Journal of Public Analysis and Management*, 18:50–76.
- Murakami, E. and J. Young, (1997) *Daily travel by persons with low income*. NPTS Symposium Paper. Bethesda.
- Pisarski, A. (1997) "Carpooling: past trends and future prospects," *Transportation Quarterly*, 51:6–8.
- Preston, V. and S. McLafferty (1999) "Spatial mismatch research in the 1990s: progress and potential," *Papers in Regional Science*, 78:387–402.
- Pucher, M. and J. Renne (2003) "Socioeconomics of urban travel: evidence from the 2001 National Household Travel Survey," *Transportation Quarterly*, 57:49–78.
- Schrank, D. and T. Lomax (2002) *The 2002 urban mobility report*. College Station: Texas A&M University, Texas Transportation Institute.
- Shen, Q. (1998) "Location characteristics of inner-city neighborhoods and employment accessibility of low-wage workers," *Environment and Planning B*, 25:345–365.
- Shen, Q. (2001) "A spatial analysis of job openings and access in a US metropolitan area," *Journal of the American Planning Association*, 67:53–68.
- US Census Bureau (2000) *Journey to work*. Washington, DC: US Census Bureau.
- US Department of Housing and Urban Development (2000) *State of the cities*. Washington, DC: Department of Housing and Urban Development.
- US Federal Highway Administration (1995) *1995 nationwide personal transportation survey*. Washington, DC: FHWA.
- US Federal Highway Administration (2001) *2001 nationwide personal transportation survey*. Washington, DC: FHWA.

*Chapter 6*

## ECONOMIC DEVELOPMENT AND TRANSPORT HUBS

KENNETH J. BUTTON

*George Mason University, Fairfax, VA*

### 1. Introduction

Transport terminals have always been a focus of economic activity. The great cities of antiquity grew up at major nodal points in trade routes – as junctions in rivers or tracks, or at natural harbors. Defense considerations and other factors acted as secondary filtering devices between the various alternative nodal points. But these cities themselves subsequently evolved and developed by acting as catalysts to draw in more economic activities and from the growth of the existing economic undertakings. This trend, albeit with clearly different nuances and often involving very different industries continues today. The aim here is to look at the role that transport hubs play in focusing economic activity at specific locations. This may be by allowing these locations to more fully reap any comparative advantage that they enjoy or by facilitating the exploitation of internal economic synergies (such as economies of scale and scope).

The links between transportation infrastructure and economic development have always been difficult to disentangle. Von Thünen's (1826) work on spatial development patterns crucially depended on transportation quality and latter supporters of the growth pole theory of development (Francois Perroux's *poles de croissance*) implicitly assumed high-quality inter-urban transportation to allow spread effects. Much of the more recent thinking has, however, been aspatial in its orientation. The macro studies of Biehl (1986), in the context of the EU, and Ashauer (1989), with its primary focus on the USA, stimulated debate about the importance of infrastructure investment, of which transportation infrastructure is a major component, on national productivity. Subsequent analysis tied this work in with more analytical modeling of economic growth and with spatial economic convergence.

The attention of this chapter is on the economic benefits that communities may enjoy through having airlines hubbing their passenger or freight services through their local airport infrastructure and, where feasible, in having international services making use of it. Similar types of analysis are applicable to other modal

hubs such as railway termini and seaports. Focusing on a particular mode has the advantage of avoiding the *ad hoc* tendency that can accompany drawing examples from very different forms of transportation.

## 2. Airports as hubs

Air transportation is not only a major industry in its own right but is of considerable significance as an input into rapidly growing national, international, and global economies. While fully accurate data are not available, air transportation accounts directly for about 1% of the GDP of the USA and of the EU. Globally, it involves a 15 million km network of services, and carries about 1500 million passengers and 30 million tonnes of cargo annually. It is now estimated that some 30–40% of world trade by value goes by air transportation. Air transportation is an essential ingredient for the success of tourism in many countries and subregions; indeed, in markets outside of Europe it is the primary mode. It is also an important input into the successful development of many, non-leisure-based industries where interpersonal communications are important. It is not only passenger air transportation that is vital to these latter industries: many such firms also rely on a range of air freight services to provide quality service to customers and to operate just-in-time production within modern chain-management frameworks.

Hub-and-spoke operations have been a traditional feature of air transportation markets. Initially this was a function of technology – early aircraft were incapable of traversing long distances without refueling, and hubbing offered an efficient solution. Later, international services were hubbed through major cities as a result of the bilateral structure of air service agreements that followed the signing of the Chicago Convention in 1944. Each country basically divided business on a route between one of its own carriers and one from another country. Traffic tended to fan out from major airports as a result. In the world's largest market, the domestic US market, the full potential of hubbing was initially thwarted by government regulation that licensed routes without due consideration of network effects.

The deregulation of the US domestic air transport market occurred in 1977, when air freight services were liberalized, and led to the very rapid adoption of hub-and-spoke operations by the larger airlines. Unlike the older systems that generally applied both in large domestic markets and internationally with the legal regime of bilateral air service agreements that had grown up after World War 2, this new hubbing structure allowed carriers to offer on-line services through hubs where a passenger did not have to change airline. Regulatory systems generally were route based and involved passengers passing through a hub to change carriers – interlining.

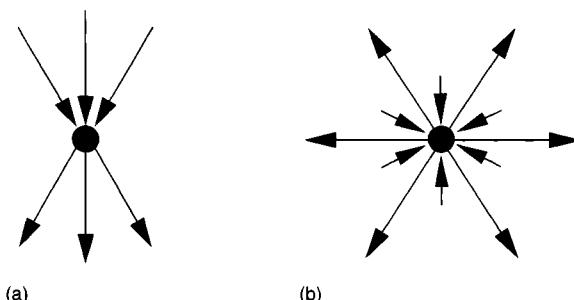


Figure 1. (a) Hourglass and (b) hinterland hubs.

These airlines were quick to exploit the potential economies of scale, scope and density that freer markets permit.<sup>a</sup> On the demand side there also came the gradual appreciation that network services provided economies of network presence that can add to the revenue flow. There are, however, limits to the economically efficient size of hub airports. As traffic grows, so does terminal and runway congestion at the airport. The congestion will tend to be concentrated at peak times as an airline coordinates flights to facilitate its on-line services.<sup>b</sup>

Airport hubs can take a variety of forms. Doganis and Dennis (1989) separate hubs into "hourglass" and "hinterland" hubs (Figure 1). The former is operated with flights from one region to points broadly in the opposite direction. A hinterland hub feeds short-haul connecting traffic to long-haul, often international, flights. The hourglass hub operation tends to use aircraft all of a similar size, whereas the hinterland hub has aircraft of mixed sizes.

With these forms of hubbing structure, flights are funneled in "banks" into a number of large hubs where substantial numbers of passengers change aircraft to complete their journeys. These banks involve the coordinated arrival of a large number of flights in a short space of time and then an equally coordinated

<sup>a</sup>These refinements on the more traditional notion of economies of scale are particularly, if not exclusively, relevant for analyzing network industries. The technical distinction between economies of scale and scope can be seen by reference to the following equation, where  $C$  denotes cost and  $Q$  is output. Economies of scope are assessed as

$$S = \{[C(Q^1) + C(Q^2)] - C(Q^1 + Q^2)\}/[C(Q^1 + Q^2)],$$

where  $C(Q^1)$  is the cost of producing  $Q^1$  units of output one alone,  $C(Q^2)$  is the cost of producing  $Q^2$  of output two alone, and  $C(Q^1 + Q^2)$  is the cost of producing  $Q^1$  plus  $Q^2$ . Economies of scope exist if  $S > 0$ . There are economies of scale if  $C/Q$  falls as  $Q$  expands.

<sup>b</sup>Congestion may be an externality if there are several carriers at an airport, but with a dominant carrier it is largely internal to that airline and reflects a conscious policy by the airline to optimize on-line connections (Mayer and Sinai, 2003).

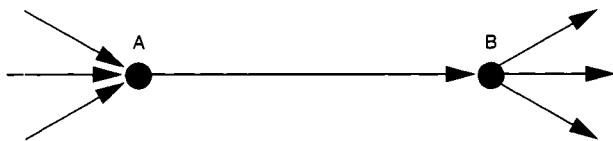


Figure 2. Multiple hub structure.

departure of flights within a narrow time window. Larger hubs may well have up to seven or more such banks a day. With the adoption of hubs, travel times were longer for many people but fares fell, and the range of potential flight combinations available to any particular destination expanded considerably.

Hubs have proved particularly important in the development of long-haul air transportation, and have served in many cases to circumvent the limitations of restrictive air service agreements. Such agreements while facilitating flights between countries have limited the ability of one country from collecting feeder traffic in another or from distribution traffic to non-gateway destinations. The growth of strategic alliances whereby airlines agree to serve as collectors and distributors of each other's traffic for cooperative international services has provided a partial solution to this. In the Star Alliance, for example, the German airline Lufthansa may consolidate European passengers at Frankfurt (A in Figure 2) and then either it or its US partner, United Airlines, flies them to Washington, DC (B). United Airlines then functions as the distributor to final US destinations. There is an increasing tendency over the denser long-haul markets such as the North Atlantic for competition between networks of these types, each involving alliances of carriers, rather than between single carriers offering point-to-point services.

Despite efforts at general categorizations, what exactly constitutes a hub airport is not clearly defined. There is no hard or fast economic or legal definition of a hub. The US General Accounting Office generally assumes a concentrated hub to be an airport that is one of the 75 busiest in the nation in terms of enplanements and where one carrier accounts for at least 60% of enplanements or two carriers combined account for at least 85%. Airports falling into either category but not in the 48 contiguous states are excluded, as are those in cities with more than one airport. The US Federal Aviation Administration classifies communities into four classes, depending on the total percentage of US passenger enplanements in all services and operated by US certified carriers within the 50 states and other designated areas (e.g. a large hub has 1.00% or more and a medium hub has 0.25–1.00%). This type of definition, however, may miss some of the key economic features of hubs, such as the degree of on-line traffic they handle.

Academic studies often look more at the degree of dominance that an airline enjoys at each airport. This may involve a simple share of flights or passengers or more sophisticated measures such as the Herfindahl-Hirschman index of

concentration. The ultimate cut-off point in the approach is arbitrary, but it does provide an indication of the extent to which carriers integrate their series. As a general rule of thumb, however, many academic studies consider a hub to entail carriers feeding three or more banks of traffic daily through an airport from some 40 or more cities. Table 1 provides a listing of the major US and European airports, and the extent to which particular carriers dominate them in terms of flights. What is notable from this is the lower degree of concentration at European airports. This is largely because they are international airports, and under bilateral air service agreements the domestic flag carrier's capacity is in general matched by that of the various bilateral partner airlines.

The benefits of hub-and-spoke operations accrue to both airlines and passengers. Airlines gain in terms of economies of scale, scope, and density on the cost side, and economies of market presence on the revenue side. Basically a hubbing airline obtains the same direct benefits and costs as a non-hub carrier but also gains additional revenue by serving routes connecting passengers. A hub airline connecting  $N$  cities serves  $N^2$  routes. If there is uniform demand across the market, each additional city added to the network increases revenue by  $2N$ . This is because passengers in the new city have  $N$  destinations available, and those in the existing  $N$  cities will have one new destination each. The residents of a hub city gain from more services, and the carriers from the ability to spread costs.

Placing exact figures on the various cost economies associated with hubbing is difficult. There are indications that due to economies of density, a 1% rise in the number of passengers an airline carries results in a 0.8% reduction in total costs, although more recent analysis indicates savings could be greater (Brueckner and Spiller, 1994).

Set against these benefits, diseconomies need to be considered. Besides potential periodic congestion, the banking of flights at hubs was traditionally seen as posing few economic problems. But it can lead to considerable periods of idle time when the number of banks per day is relatively small. Ground staff and other resources are left with little to do, and aircraft are used much less effectively. The optimal degree of banking is when the marginal benefit from reducing connecting times is equated to the additional congestion costs of concentrating flights in narrow time widows, including the cost of down-time between banks when the gates and slots of an airline are left idle. Hubbing can also mean a spatial concentration of traffic with environmental intrusion for those living in the region. Airports are noisy, can lead to soil contamination from run-offs of, for example, chemicals in de-icing fluids, and often create congestion in the local surface transport network.

From the passenger or freight consignors' perspective there may be concern that hubbing confers monopoly power on the major carrier at an airport. The carrier is in a position to charge high fares to those captive to that airport. In particular, since business travelers are generally less sensitive to fare levels (i.e.

**Table 1**  
Flights by the three lead carriers at US and European airports (May 2003)

Airlines and their proportion of flights from each airport			
Airport (a)	Carrier 1	Carrier 2	Carrier 3
<b>Top 10 US airports by passenger ranking</b>			
Atlanta Hartsfield (1)	Delta Air Lines, 73.7%	AirTran Airways, 14.6%	American Airlines, 2.3%
Chicago O'Hare (2)	United Airlines, 47.0%	American Airlines, 38.6%	Delta Air Lines, 2.2%
Los Angeles (5)	United Airlines, 30.8%	American Airlines, 2.3%	Southwest Airlines, 13.8%
Dallas/Fort Worth (6)	American Airlines, 63.2%	Delta Air Lines, 19.0%	United Airlines, 1.6%
Denver (10)	United Airlines, 53.3%	Frontier Airlines, 15.1%	Great lakes Aviation, 12.2%
Phoenix Sky Harbor (11)	America West Airlines, 51.1%	Southwest Airlines, 27.7%	United Airlines, 3.6%
Las Vegas McCarran (12)	Southwest Airlines, 39.6%	America West Airlines, 20.3%	United Airlines, 8.7%
Houston George Bush (13)	Continental Airlines, 82.7%	American Airlines, 3.4%	Delta Air Lines, 3.0%
Minneapolis (16)	Northwest Airlines, 80.3%	American Airlines, 3.6%	Delta Air Lines, 2.9%
Detroit Wayne County (17)	Northwest Airlines, 79.4%	American Airlines, 3.8%	Delta Air Lines, 2.8%
<b>Top 10 European airports by passenger ranking</b>			
London Heathrow (3)	British Airways, 41.6%	BMI, 12.1%	Lufthansa, 4.8%
Frankfurt (7)	Lufthansa, 59.4%	British Airways, 3.6%	Austrian, 2.9%
Paris Charles de Gaulle (8)	Air France, 56.6%	British Airways, 5.1%	Lufthansa, 4.9%
Amsterdam (9)	KLM, 52.2%	Transavia, 5.5%	easyJet, 4.3%
Madrid (14)	Iberia, 57.0%	Spanair, 12.7%	Air Europa, 7.1%
London Gatwick (21)	British Airways, 55.1%	easyJet, 12.8%	flybe British European, 5.6%
Rome (31)	Alitalia, 46.2%	Air One, 10.0%	Meridiana, 3.9%
Munich (35)	Lufthansa, 56.8%	Beutsche BA, 6.6%	Air Dolomiti, 6.5%
Paris Orly (36)	Air France, 64.2%	Iberia, 8.2%	Air Ittoral, 3.6%
Barcelona (38)	Iberia, 48.55	Spanair, 9.4%	Air Europa, 5.5%

Source: *Airline Business*, June 2003.

Note: (a) world ranking by passengers in parentheses.

less price-elastic) because of their wider view of the generalized costs of making any trip, this group is potentially a soft target for hub airlines. The effect could, therefore, be negative for local business development. The development concern, however, is not with narrow fare comparisons between regions with and without hubs but rather with the overall effects of having a hub airport located in a region. However, business travelers are generally much less price-sensitive but exhibit more demands on quality of service (e.g. in terms of time and frequency of flights and destinations served directly). Survey evidence indicates, for example, that schedule convenience (especially frequency) is by far the most important factor for business travelers' choices of airline and is the second most important feature for leisure travelers (Ostrowski and O'Brien, 1991). By offering flexible tickets, comfortable on- and off-plane amenities, and scheduling convenient-to-business needs at a competitive price, carriers can attract these users.

### **3. Regional impacts of hub airport**

Airports have four potential impacts on the economy in their region.

#### *3.1. Primary effects*

Primary effects are the benefits accruing to the region from the construction or expansion of the facility – the design of the facility, the building of the runways, the construction of the terminals and hangers, the installation of air traffic navigation systems, and so on. The direct effects of this involve the local employment required in the construction process and the work done by local contractors. Indirect effects include the benefits to the region of the wages and other incomes that these workers and companies subsequently spend in the area. These are clear gains to the local economy, but they are short-term, once-for-all, and may be rather small. Also, airport construction involves a degree of specialist skill, personnel, and equipment that may not be available locally, and this leads to leakage. In general, while airport development can have beneficial primary effects, save in cases where there is a policy imperative to create jobs in the very short term, these are not really the key concerns.

#### *3.2. Income multiplier effects*

Income multiplier effects are longer-term and are associated with the local economic benefits of running and operating the airport – the employment in maintaining the facility, in handling the aircraft and passengers, in transporting

people and cargo to and from the terminal, and so on. Again there are direct effects stemming from the immediate jobs that are created at the airport and immediately associated with it. There are also indirect effects due to the ongoing flow of income that the airport's operation puts into the local economy. These secondary effects can be extremely important to a local economy in terms of employment, income, and, for local government, taxation revenue. The actual size of the multiplier effect will vary between airports, dependent upon the nature of their operations.

### *3.3. Tertiary effects*

Tertiary effects, which are not strictly Keynesian multipliers although often called such, stem from the stimulus enjoyed by a local economy as the result of firms and individuals having an extensive system of direct air transport services at their disposal. Typical hub city air services offer:

- more frequent flights;
- more direct flights;
- more opportunities for same-day return flights;
- greater likelihood of international flights;
- services geared to local market needs;
- the ability to send packages on scheduled passenger services on flights leaving after the major courier services have finished their daily pick-ups;
- residents of hub cities the same opportunities of linking to other major hubs as those living in non-hubs.

These may be seen as important features for business travelers. The cost of the average business trip is not assessed purely in terms of air fares. Generalized costs that embrace, among other things, air travel time, time spent in terminals, time spent getting to and from airports, costs of getting to and from airports, costs of overnight stays, and costs of time wasted due to infrequent flights are often more important.

### *3.4. Perpetuity effects*

Perpetuity effects reflect the idea of "new growth theory" that economic growth, once started in a region, becomes self-sustaining and may accelerate. Linked to this, there is empirical evidence that infrastructure investment can act as a catalyst for higher economic growth in an area; essentially it can act as a kick-start mechanism. The construction of a new airport may set in progress a larger and longer-term development process in a region. This perpetuity effect is in addition

to the tertiary effects that relate to the immediate migration of firms to an area with good air transport services. It is longer-term and affects the dynamics of an area. By initially attracting undertakings in sufficient numbers, airport development can lead to the crossing of important thresholds in terms of economies of scale, scope, and density. In particular, in the context of “new economy,” high-technology activities, an area can acquire a vital knowledge base that fosters local research and development and makes the region quasi-independent of others.

This type of dynamic economic impact of an airport is the most abstract and the most difficult to quantify. It has been little researched and involves the interaction of an airport with all other aspects of the local economy. It is, nevertheless, potentially a very real and important benefit that may be enjoyed by a region with high-quality air services. It is difficult to see, for example, how numerous Mediterranean and Caribbean islands could have been transformed from fishing communities to major tourist centers without the creation of airports. In the high-technology context, the M4 motorway corridor in the UK was strongly tied to the global access offered by Heathrow Airport, and the Northern Virginian high-technology region linked to the access offered by Dulles Airport.

#### 4. Empirical analysis

Efforts to quantify the local economic implications of hub airports are as much a practical expediency as an academic exercise. Many countries require economic impact assessments before the development of an airport or the significant expansion of an existing one. Ideally, from a national income perspective these should be done within a general equilibrium framework to allow for crowding out effects. In practice they often take more of the form of a spatial impact study focusing on the area immediately adjacent to the airport location. A number of approaches are used. Box 1 provides a summary of types of magnitude that have emerged.

##### 4.1. Surveys and questionnaires

This approach has its background in market research, and is particularly useful for some types of forecasting. It is helpful, for example, when seeking information about entirely new facilities or airline “products” that is not easily obtained by studying current behavior patterns.

At one level the technique involves asking affected parties about the role that an airport plays in their commercial decision-making (e.g. regarding location, markets served, and scale of activities) and the air travel that is undertaken. More advanced techniques through the application of stated-preference methods deploy carefully

**Box 1**  
**Examples of airport impact studies**

**Survey techniques**

- The Atlanta Chamber of Commerce found from a survey of 264 foreign-based firms that the availability of direct international services was the third most important factor in location decisions. A subsequent study showed that the number of foreign firms locating in the region from a particular country grew significantly after the introduction of a non-stop service.
- Ernst and Young, looking at location decisions of 57 companies in Europe making decisions regarding the location of a manufacturing plant, found that the air transport network was the third most important factor in the decision process. Air services were much more important for service sector companies.
- The Amsterdam Chamber of Commerce found that the availability of an airport was one of five key factors considered in company relocation decisions.
- A survey of firms around Zurich found that 34% and 38% considered the airport to be very important and important, respectively, as a location factor.
- Loudoun Chamber of Commerce (Virginia) found that airport/freeway access was important to 68% of firms.
- A study of small business firms in Washington, DC, that were engaged in export activities found the availability of easy access to international air transport one of the six most important factors in their success.

**Multiplier techniques**

- An academic study by Rietveld estimated that Schiphol Airport (Netherlands) generates about 85 000 jobs for the country.
- A study of Vienna International Airport by the Industriewissenschaftliches Institute in 1998 indicated that on a turnover of ATS 25 billion in 1996, there was an impact of ATS 11.2 billion on the local economy.
- The Institute of Social and Economic Research found that the total annual economic importance of Anchorage International Airport on local payrolls was US \$130 million above the US \$316 million for on-site activities.

**Econometric techniques**

- Analysis by George Mason University taking variations in high-technology employment across all US metropolitan standard areas, found that a hub airport in a region increases that region's new economy employment by over 12 000.
- Brueckner, looking at possible expansion of Chicago O'Hare Airport, found that an increase of traffic of 50% will increase service-related employment in the region by 185 000 jobs.
- An econometric study by Science Applications International of the implications of the Open Skies agreement between Germany and the USA on the regional economy around Hamburg Airport found annual gains for the regional economy of US \$783 318 (1994 value).
- Button and Taylor examined the impact on US cities of having European services, and found that employment was systematically positively related to both the scale of services offered by airports and the range of destinations served.

Source: Button (2004).

structured questionnaires related to a series of scenarios. This helps ensure that those questioned have a common perspective regarding background conditions and the options available. These can provide information about what residents want from their local airport in the context of their willingness-to-pay. Comparisons can then be made with what is actually taking place.

Designing the questionnaire is a difficult task, and there are problems in selecting an appropriate sample. There is always an inherent danger in these stated-preference studies that respondents will try to manipulate their answers to their advantage rather than give a genuine reply. For example, businesses may try to manipulate their responses to obtain lower fares but at the same time try to get better services at the expense of other air-traveling groups.

In practice, there is often a tendency by lobbying groups to use such methods to support a particular position. While there are carefully crafted academic studies of the local implications of an airport, there are far more that have been conducted by groups with a vested interest in the outcome of the decision process.

#### 4.2. Multiplier analysis

Most economic impact studies have made use of some form of standardized analysis, sometimes linked with an input-output table, to assess the primary and multiplier effects of an airport. In many instances such impact analysis is a statutory requirement if an airport is to be built or an existing one is to be modified. The techniques sometimes seek out local leakages to an impact multiplier but more frequently make use of national parameters or those of studies conducted elsewhere. Taking a fairly standard formulation of a local impact multiplier highlights problems in such an approach.<sup>a</sup> Some are specific to airport investments, others more general. Many of the initial inputs into a modern airport, and especially the electronics, are seldom produced at the site of the airport or in the adjacent region. This immediately reduces the size of the multiplicand. The taxation and import leakages from a region are very difficult to assess, and the inherent problems extend well beyond assessing airports.

Setting aside the inherent problems of the Keynesian underpinnings of this approach and the difficulties of gathering accurate data to calculate parameters, they are also sometimes misused. In particular, the technique is extended to embrace tertiary effects. The vast majority of studies of this type overestimate the impacts of air services for two reasons (Butler and Kiernan, 1992). First, most studies assume that visitors to a location will fall to zero without the air services under analysis; the assumption is one of perfectly elastic demand with respect to the relevant travel determinants. In practice, there are alternative modes, and often alternative airports, that will bring some of these travelers to the region.

<sup>a</sup>At its simplest this is  $[I - M]/(mps + mpt + mpm)$ , where  $I$  is the initial investment in the local airport,  $M$  is the amount of this multiplicand that has to be immediately imported from another region,  $mps$  is the marginal propensity to save,  $mpt$  the marginal propensity for taxation, and  $mpm$  is the marginal propensity to import from outside of the region.

Second, air services may bring in travelers but they also take residents out. This will, for example, reduce the net gains from tourism as some residents will take vacations elsewhere and some local businesses will locate some of their activities in other regions.

#### *4.3. Econometric models*

Here, an econometric approach is favored that compares the economic performance of cities that have hub airports with those that do not. This macro-level analysis can offer useful insights but it also has its limitations. In particular, there are many things that influence the economic performance of a city other than simply the nature of its links with the scheduled air transport system. It is possible, however, by using standard statistical procedures, and in particular hedonic indices, to make adjustments for many of these factors and to tease out some information concerning the role of hubs.<sup>a</sup> In doing this it is helpful to pay particular attention to those aspects of cities' economies that are likely to be most tightly tied to the availability of high-quality scheduled air services.

A problem with hedonic indices is that of defining and quantifying the factors that go into a comprehensive hedonic index specification. Regional development is difficult to explain, and attempting to specify a model that contains adequate *ceteris paribus* conditions that allow the peculiar effects of air transportation to be extracted is a large task. Linked to this is the technical issue of multicollinearity, which can be serious when a large number of variables are fed into a regression model of this type.

A second problem of specifying an econometric model in this field is that of causality. Fully specified systems are seldom possible because of their complexity and data limitations. Reduced-form models must implicitly assume a direction of causality from the availability of a hub airport to local economic development. An alternative view is that a rapidly growing area will stimulate the creation of hubbed air services because demand is high and the revenue potential for airlines justifies the supplying of services. There have been few tests of causality. However, the limited work that has been done does seem to indicate that it is from the airport hub to local economic development rather than vice versa.

<sup>a</sup>A hedonic index approach takes the general form  $E_j = F(A_j, A_i)$ , where  $E_j$  is the employment or income in region  $j$ ,  $A_j$  is a vector of the attributes of region  $j$  (including the availability of air services), and  $A_i$  is a matrix of the attributes of other regions  $i$ . The parameters associated with the coefficient of  $A_j$  relating to air services provides an indicator of the importance of this feature to the state of the region's economy.

**Table 2**  
Jobs and income from each additional million passengers using an airport

Level	Jobs		Economic Impact (US \$ millions)	
	Direct	Total	Direct	Total
High	2000	8000	225	1600
Medium	1500	6000	75	650
Low	750	2500	35	130

Source: Air Transport Action Group (2000).

As an example, in a study of US airports from 1979 to 1997, Button and Lall (1999) used Granger causality tests<sup>a</sup> to see if the growth in high-technology employment in Cincinnati (a Delta Airlines hub) and Pittsburgh (a US Air hub) was caused by the availability of air services or acted as a catalyst for airlines to move into their metropolitan regions. The tests indicated that at the 95% confidence level the passenger traffic going to these airports created positive local employment effects rather than being attracted by prior demand.

From a practical perspective it is often useful to have a broad guide to the impacts on local economies of the availability of air services. Some indication is offered by data supplied by the Air Transport Action Group (2000), and is set out in Table 2. These data are an indicator of the potential jobs and income effects (effectively the multiplier effects and the tertiary effects) of each million passengers using an airport; however, they should be taken as maximum figures since their source is essentially a lobby group, but they form a useful benchmark for analysis.

## 5. International airline hubs

There is a powerful ongoing trend to liberalize the international air transportation market. The free market within the EU is the most pronounced manifestation of this, but the US Open Skies policy has led to over 50 liberal bilateral agreements between it and other countries, and there are multilateral agreements emerging in Africa (e.g. the Yamoussoukro Decision) and in the Pacific area. Given that this

<sup>a</sup>It is important to note that the statement  $x$  Granger causes  $y$  does not imply that  $y$  is the effect or the result of  $x$ . Granger causality measures precedence and information content but does not by itself indicate causality in the more common use of the term.

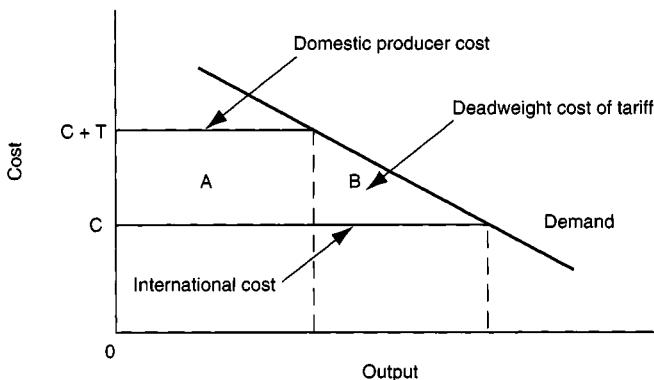


Figure 3. The benefits of freer international air services.

trend is paralleling wider moves favoring free trade, there are arguments that hub airports with international services offer additional regional economic benefits.

The economic case for removing restrictions on international air transportation is identical to those that emanate from freer trade *per se*. The arguments are not new and are some of the most robust in modern economics. At the simplest level, freer trade in any activity, including air transportation, allows customers more choice and ensures that the most efficient producers deliver goods and services. The argument can be seen in diagrammatic terms in Figure 3.

The efficient price for international air transportation established in a competitive, international environment would be  $C$  – the natural, competition-based international cost of buying the service. The constraints imposed by restrictive bilateral air service agreements, regulations over foreign ownership, the lack of cabotage rights, and so forth, however, add to this cost. The result is that consumers of air transport are confronted with prices of  $(C + T)$ . The result is that fewer services are consumed and that fares are higher. There is an overall loss to society equal to the triangular area B in the diagram. While area A is not in economic terms a loss but is seen as a transfer of benefits from consumers (who have to pay higher fares under the restrictive regime) to airlines, some of this cost may also reflect lower efficiency on the part of protected carriers.

Problems emerge in trying to place numbers on these various elements. The challenge is to generate credible estimates of the likely source, magnitude, and distribution of incremental costs and benefits that would flow from negotiated changes in bilateral agreements or their removal. Some of these benefits are, in principle, directly quantifiable; for example, improvements in economic efficiency, fiscal objectives (tax revenues and expenditures, trade exports and imports, and the balance of trade), employment objectives (number and quality of

jobs, and development of tourism and related industries), and the cost of travel to the business and tourist client. The difficulty with itemizing effects is that it is virtually impossible to avoid double counting and with it the potential of overestimating the gains from liberalizing the international institutions. Taking individual case studies of airports and then applying conventional multipliers to gain insights into the employment effects of additional international services is another alternative. The problem is that of determining appropriate multipliers for each airport and in holding other, non-air transportation effects constant.

One method of trying to capture some of these effects is by looking at fare changes. The initial thrust of the Open Skies policy in 1979 and prior *de facto* initiatives in stimulating the introduction of more liberal bilateral agreements by the USA between 1976 and 1981 were estimated by Dresner and Tretheway (1992) to have generated as much as US \$325 million savings in North Atlantic fares alone in 1981. There are severe data problems in this approach. More global analysis was limited by lack of complete data from less developed countries. Even the routes that were included could only be assessed in terms of either the full economy fare or the lowest minimum fare. In addition, it ignores what economists call the consumer surplus benefits that exceed fare savings. These embody the difference between how much individuals would have been willing to pay for a flight and what they actually pay following market liberalization.

While measuring economic performance at the national level is difficult, it poses even greater problems at lower levels of spatial disaggregation. Again the focus is on more mobile industry and its location. A limited amount of prior empirical work has been conducted on the links between new economy employment and the availability of international air transportation services. Much of this has involved interviews with local business. The findings generally indicate that these types of service can be influential in affecting the location decisions of employers. An early US study by the Atlanta Chamber of Commerce in 1988, for example, involving the survey of 264 foreign-based establishments, found that the availability of direct international services was the third most important factor in location decisions by these firms. A subsequent study by the Atlanta Chamber of Commerce showed that the number of foreign firms locating in the region from a particular country grew significantly after the introduction of a non-stop service.

The types of findings are fairly consistent across countries (see again Box 1). For example, work by Ernst and Young, looking at the location decisions of 57 companies in Europe making decisions regarding the location of a manufacturing plant, found that the air transportation network was the third most important factor in the decision process. The study concluded, however, that air services were much more important for service sector companies. Similarly, the Amsterdam Chamber of Commerce found that the availability of an airport was one of five key factors considered in company relocation decisions. A survey of firms around Zurich found that 34% and 38% considered the airport

to be very important and important, respectively, as a location factor. The Loudoun Chamber of Commerce, in Virginia, found that airport/freeway access was important to 68% of firms. A study of small business firms in the Washington area that were engaged in export activities found the availability of easy access to international air transportation one of the six most important factors in their success (Lin Salem, 2000).

Numerical methods have focused purely on the more direct effects of air services on a local economy, namely those associated with the airline service *per se* and the multiplier implications of this for the local economy. This was the basic approach underlying a study looking at ten US airports prepared for USA-BIAS (Kurth, 1990). It looked at the implications of a service to Tokyo and one to London. The latter, involving a Boeing 767 operation carrying 100 000 passengers per year, was calculated to involve direct annual expenditure by the airline (excluding fuel) of US \$2.9 million (1990 prices) and expenditure by incoming tourists of US \$30.7 million. Applying parameters from an earlier study of Houston-Tokyo services indicate that a service to London would generate an annual increase in exports of US \$84 million (University of Houston, 1988).

A study by the UK Civil Aviation Authority (1994) looked at the importance of new international services involving airports outside of London. In terms of North Atlantic services, it considered the economic implications of new services between Birmingham and New York, and between Manchester and Atlanta. The study essentially used multipliers, and found that, after allowing for traffic diversion effects as well as traffic creation, the New York route would generate some UK £1 million in passenger benefits per annum, and the Atlanta service UK £1.2 million. A more rigorous econometric study of the implications of the Open Skies agreement between Germany and the USA on the regional economy around Hamburg Airport found that labor demand at the airport went up by 19 people due to increased passenger demand, and by 28 due to increased flights. A further 78 jobs were generated in the region due to the availability of enhanced transatlantic air services. In value terms this amounted to an annual gain for the regional economy of US \$783 318 in 1994 prices (Science Applications International et al., 2000).

To assess the importance of international air access to regions of the US economy, Button and Taylor (2000) examined the benefits of US cities having European air services. The analysis makes use of US Department of Transportation, Bureau of Transportation Statistics, data. It takes 41 US airports of differing size and from various parts of the country. The airports were selected as large and medium-sized airports. The larger airports are included to allow the hypothesis regarding the diminishing marginal gains from additional international air transportation services to be tested. Most of the airports offer international services, although not all to Europe (or to the EU area). The aggregate analysis is largely based on the entire set of airports, but the subsequent case study analysis

looks only at a subset to allow a more detailed evaluation through time. The regions that the airports serve extend across the USA, and represent a range of geographical types with diversity in their economic compositions.

The method of analysis was least-squares multiple regression, which was used to determine parameters for a non-linear model linking new economy employment in a region to a set of parameters including the availability of European air services and their utilization. Europe in this context is taken to be the 15 EU nations together with Norway and Switzerland. Information regarding other countries is available, but these 17 nations account for the majority of transatlantic traffic and provide a significant business base that allows interaction with the higher-technology companies in the USA.

The effects of international air transportation on the dispersion of new economy employment are captured in two ways: the scale of services offered by airports, and the range of destinations served. The number of destinations served is included directly, and the scale of service is proxied by onboard passengers. Since there is an inevitable time lag between the provision of a new air service and its full integration into the economic parameters of a region, the analysis assumes a 2 year adjustment period. Hence, the employment structure in each region in 1996 is seen as influenced by, among other factors, the international air transportation services available in 1994. Further, since the influence of the quality of European air services is likely to decline the further west a US city is located, an attempt to capture this distance effect is made by including a time zone variable. One would anticipate that this should take a negative sign.

The analysis incorporated a range of other variables that seemed intuitively important and have sometimes been found to be significant in other studies of economic development. Since there is evidence from the airline hubbing analysis that high-quality air services are important *per se*, the analysis includes the total number of enplanements to reflect this and to capture the importance of domestic services as well as European international services. The resultant model explained over 80% of the overall variation in new economy employment across the 41 US metropolitan statistical areas that were included in the analysis.

In the context of the availability of air transportation, three variables all have positive coefficients. A total enplanements variable is small, but this may be expected since much of the air transportation effect is taken up by the international dimension of the equation. There is a tendency for larger airports to have more international activities. With regard to international air services, the associated coefficients for both on-plane passengers and the number of destinations served take positive signs. They exert a positive influence on the attractiveness of regions to new economy employers. As important for confirming the robustness of the specification is the fact that variables such as the overall level of air service, population, and military employment are positive, as would be anticipated.

What these coefficients mean in qualitative terms depends upon the nature of changes in international air services. Often of interest is the issue of changing the scale of international services at an airport already having a portfolio of international destinations to serve. But it would normally be incorrect to simply take the introduction of an additional destination without allowing for the additional capacity (proxied by on-board passengers) that inevitably comes with it. As an example, an increase in EU destinations served from three to four in itself means about 1760 extra new economy jobs for a region. But an extra destination would almost inevitably mean additional capacity. A reasonable assumption is that the number of on-board passengers increases from about 120 000 to 160 000. The impact of this on new economy employment is an additional 1150 new economy jobs in the region. Hence the total employment effect for the regional economy is about 2900. If each job is assumed to be worth US \$55 000 per annum (this includes salary and employer contributions to health plans, pensions, etc.) this is a US \$160 million per annum benefit. Discounted at 5% per annum, this over 10 years represents a capitalized value of over US \$1235 million at constant prices.

## 6. Conclusions

This chapter has examined some aspects of the links between the ways in which airports are used and local economic development in the USA. It is found that growth in the new economy sectors in a region are positively affected by the degree to which the local airport serves a major hub function in the domestic market and has strong international ties. In some instances the findings are seen as clarifying what have often been largely anecdotal debates, e.g. over whether any hub premium acts to deter local investments in industry. In others it provides quantification of some of these effects that have been widely accepted but have to date only been the subject of cursory quantitative analysis, e.g. the impacts of having differing levels of international air access for different regions.

The findings provide support for Winston (1991) and others who have criticized the work of Aschauer (1989) and Biehl (1986) for failing, among other things, to look at the use made of infrastructure as well as at aggregate investment levels. The results, because of their comparative nature, should, however, be seen as only indicative of the role that various uses of air transportation infrastructure play in regional development. They reflect comparative situations, and it is quite possible that all the observation points (e.g. relating to the number of international services) are within the potential production possibility frontier. Indeed, given the institutional constraints still present in air transportation this seems inevitably to be the case. They do, nevertheless, indicate that hubs are not the ogre of local

industrial growth that has sometimes been suggested, and provide further support for the benefits of more liberalized air transportation markets.

## References

- Air Transport Action Group (2000) *The economic benefits of air transport*. Geneva: ATAG.
- Aschauer, D.A. (1989) "Is public expenditure productive?" *Journal of Monetary Economics*, 23:177–200.
- Biehl, D. (1986) *The contribution of infrastructure to regional development*. Brussels: Regional Policy Division, European Communities.
- Brueckner, J.J. and P.T. Spiller (1994) "Economics of density in the deregulated airline industry," *Journal of Law and Economics*, 37:379–415.
- Butler, S.E. and L.J. Kiernan (1992) *Measuring the regional economic significance of airports*. Washington, DC: Office of Airport Planning and Programming, Federal Aviation Administration.
- Button, K.J. (2004) *Towards an efficient European air transport system*. Aldershot: Ashgate.
- Button, K.J. and S. Lall (1999) "The economic of being an airline hub city," *Research in Transport Economics*, 5:75–106.
- Button, K.J. and S.Y. Taylor (2000) "International air transportation and economic development," *Journal of Air Transport Management*, 6:209–222.
- Doganis, R. and N.P.S. Dennis (1989) "Lessons in hubbing," *Airline Business*, March:42–45.
- Dresner, M. and M.W. Tretheway (1992) "Modeling and testing the effect of market structure on price: the case of international air transport," *Journal of Transport Economics and Policy*, 25:171–83.
- Kurth (1990) *Better international air service: economic base and economic benefits*, Washington, DC: Kurth.
- Lin Salem, P. (2000) "Local dynamics and service exports; an analysis of the internationalization of small business service firms in the Washington, DC metropolitan area," Ph.D. dissertation. Fairfax: George Mason University.
- Mayer, C. and T. Sinai (2003) "Network effects, congestion externalities, and air traffic delays: or why not all delays are evil," *American Economic Review*, 93:1194–1215.
- Ostrowski, P.L. and T.V. O'Brien (1991) *Predicting consumer loyalty in for airline passengers*. DeKalb: Department of Marketing, Northern Illinois University.
- Science Applications International, MKmetric and Institut für Weltwirtschaft (2000) *The Impact of liberalizing international aviation bilaterals on the northern German region*. Hamburg: Science Applications International, MKmetric and Institut für Weltwirtschaft.
- UK Civil Aviation Authority (1994) *The economic impact of new air services: a study of long haul services at UK regional airports*, CAP 638. London: CAA.
- University of Houston (1998) *The impact of Tokyo direct flights on the Houston economy*. Houston: University of Houston.
- von Thünen, J.H. (1826) *Der Isoline Staat in Beziehung auf Nationale-Konomie und Landwirtschaft*. Stuttgart: Gustav Fischer (1966 reprint).
- Winston, C. (1991) "Efficient transportation infrastructure policy," *Journal of Economic Perspectives*, 5:113–128.

*Chapter 7*

## TRANSPORT AND SPATIAL CLUSTERING

JEAN H.P. PAELINCK

*George Mason University, Fairfax, VA*

### 1. Introduction

Local concentrations of economic activities have interested spatial – and even general – economists for many decades; an early example is Alfred Marshall in his *Principles of Economics*, an aspect of which has recently been “rediscovered” (e.g. see Akoorie, in Green and McNaughton, 2000). Contemporary economists and geographers still keenly share that interest.

The presentation is organized in the following manner. My introductory section concerns terminology, which is far from standard. The concepts it aims at clarifying are often loosely defined. I propose a nomenclature in line with rigorous definitions of the phenomena to be described and analysed.

The traditional market factors (prices and quantities) approach then serves as a starting point for further developments. A clear example of its use is the way in which haulage and transport costs play a central role in the so-called Tinbergen–Bos systems (TBS) approach, and gives a particular insight into the problem at hand. Specifically it generates conditions under which local concentrations of activities will be observed – or not – and also the types of joint locations that result from those activities.

More recently, other, non-market, factors have been brought to the fore, competing with – or complementing – the traditional factors that have guided the explanation of the joint location of economic activities. A central point is the transportation of goods versus mobility of people and ideas where an investigation into the assertion of declining transport costs is in order.

The final two sections provide, first, a synthesis of the main results in the field, then conclusions as to what the future is likely to bring in terms of facts and analyses.

A limited number of references are provided as a starting point for further information and study.

## 2. Preliminary concepts

As said earlier, the terminology is far from uniform, which is essentially due to the fact that little attention has been given to rigorous definitions of what is meant by local concentrations of economic activity; hence the following developments. Casual observation already shows that economic activity is unequally dispersed over geographical space; the task then is to develop and define clear concepts that could be used in further analysis.

One needs in the first place the notion of a distance or metric; this being given, consider any local presence of economic activities.

A cluster can then be defined in the following way: it is a set of localized economic units (production plants, service outlets, etc.) with subscripts  $k$  and  $l$  that obey the condition

$$\forall k \exists l \mid d_{kl} \leq d^*, \quad (1)$$

or in plain words: for every activity unit  $k$  one can find one or more activity units  $l$  that lie within a distance  $d^*$  from  $k$ . A cluster is thus a purely geographical concept (*a contrario* Czamanski, 1974) that could be applied to any type of elementary object. An immediate consequence of this definition is that the number of clusters over a given reference area, a country say, will increase with smaller  $d^*$ ; Figure 1 illustrates this fact.

If  $d^*$  is relatively small, three clusters will appear; for large  $d^*$ , only one cluster will be present. Also, if all units have the same value – whatever the measure used – with decreasing  $d^*$  the local density, defined as the number of units over the convex hull of the cluster, i.e. the smallest convex figure that contains the cluster, will decrease with respect to the highest density in the previous partial clusters.

The next concept needs the introduction of a techno-economic magnitude, to wit, an input coefficient. The latter is defined as the share of a product or service, used in the production of a given item, in that item's total production value, and is generally noted  $a_{kl}$ ,  $k$  being input to  $l$ . This allows the definition of a complex as a cluster, which obeys the following conditions:

$$a_{kl} * a_{lm} * a_{mn} * \dots * a_{st} * a_{tk} \neq 0, \quad (2)$$

meaning that there exists an uninterrupted closed chain of input relations between

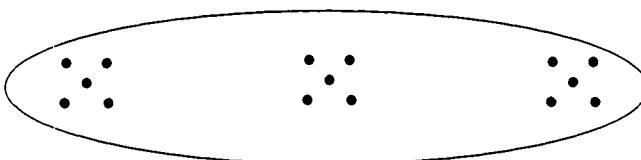


Figure 1. Clusters.

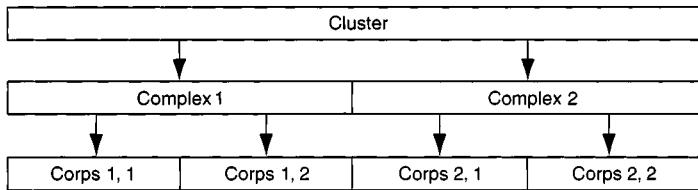


Figure 2. Cluster, complexes, and corps.

a set of clustered activities. In mathematical terms, the corresponding input matrix is not decomposable, which implies that any shock, imposed on one of the activities, will transmit itself to all the other ones. One can sharpen the definition by imposing  $a_{kl} \geq a^*, \forall k, l$ , which will reduce the size of the complex; some of these complexes can be rather large: in the Washington, DC, area, for example, a complex of 371 activities (and many more production units, of course) has been detected, out of 383 relevant activities (Stough et al., 2002).

A final notion is that of a corps. It is defined as a sub-complex that maximizes a partial sum of the relevant interactivity flows, still leading up to a non-decomposable set of activities. In that sense a complex can be made up from several corps, possibly in declining order of internal cohesion, measured by the above-mentioned sum.

Figure 2 presents the hierarchy of the three concepts, as a background for the 3C-analysis, which follows.

As already said, the terminology developed above is not currently in common use in the literature – most authors use “cluster” where here “complex” is proposed – but the definitions advocated here tie directly to the mathematical literature on cluster analysis.

The next section will present the traditional approach to spatial activity concentration.

### 3. Market approach<sup>a</sup>

Cluster and complex analysis is examined here from the perspective of an approach developed in the 1960s by Jan Tinbergen and Henk Bos; it was designed to derive propositions on economic “landscapes” in terms of clusters of activities (“centers”) – possibly complexes, as will be seen – with a specific combination of clusters or complexes being called a “system” (for further details see Paelinck, 2001).

<sup>a</sup>Adapted from Paelinck (2004).

The “system” – in fact a spatial economic equilibrium – is computed by minimizing, for given unit prices, an objective function including consumption (“propensities”) and production (“input”) coefficients. The composition and location of “centers” – the “economic landscape” or “system” – are then determined by the minimization of global transport costs (though this may correspond to profit maximization behavior by individual firms), i.e. of a function  $\varphi = w \times x$ , where  $x$  is a series (“vector”) of exports of goods and services, and  $w$  is a vector that depends upon the distances  $d_{ij}$  between all the possible locations  $i$  and  $j$ , on unit transport costs, and on relative quantities shipped, the latter in turn depending on the consumption propensities and technology coefficients defined as follows:

- $a_k$  is the propensity to consume product  $k$ ;
- $y^*$  is the total value of production (or value added in the absence of inter-industry relations) of the system;
- $a_k y^*$  is the production of sector  $k$ .

So, in the absence of inter-industry relations, the value transported between sectors  $k$  and  $l$  equals  $a_k a_l y^*$ . Each firm of sector  $l$  demands  $a_k a_l y^*/n_l$  from sector  $k$ ; there are  $n_l$  plants of type  $l$ . The total weight for deliveries between  $k$  and  $l$  thus becomes  $(t_k + t_l)a_k a_l y^*$ , so that the relative weights (excluding distances) relating to the above mentioned flows become  $w_{kl} = (t_k + t_l)a_k a_l n_l^{-1}$ , where  $t_k$  and  $t_l$  are the unit transportation costs.

If  $a_{kl}$  now is the input coefficient of  $k$  in sector  $l$ , and  $m_k$  is the production versus value-added ratio of sector  $k$ , then for deliveries between sectors  $k$  and  $l$  the total transportation costs become

$$t_{kl} = [t_k(a_k + a_{kl}m_l)a_l + t_l(a_l + a_{lk}m_k)a_k]y^*. \quad (3)$$

How complicated such interactions can be is shown by Figure 3, which pictures the potential flows between one agriculture sector, sector 0, and two industrial sectors, sectors 1 and 2 (indices of the flows; the indices inside the squares refer to their spatial characteristics – agricultural activities 0, centers producing goods 1 or 2, and center 3 including both types of production). Final flows go to consumption; intermediate ones are interactivity deliveries.

Figure 4 shows how varying elements of function 3 can affect the economic landscape; the relative weights,  $w_3/w_1$  refer to deliveries between industries 1 and 2, on the one hand, and activities 1 and 0 (agriculture) on the other. The causes of those changes may be multiple: changes in transportation costs, in consumption propensities, or in technological data. The crosses refer to activity 1; the circles to activity 2.

In fact, the spatial concentrations generated might be clusters, complexes, or corps; looking at function 3 only one-way input-output relations could be present, or they may be circular, and, inside the latter complexes, corps might be present.

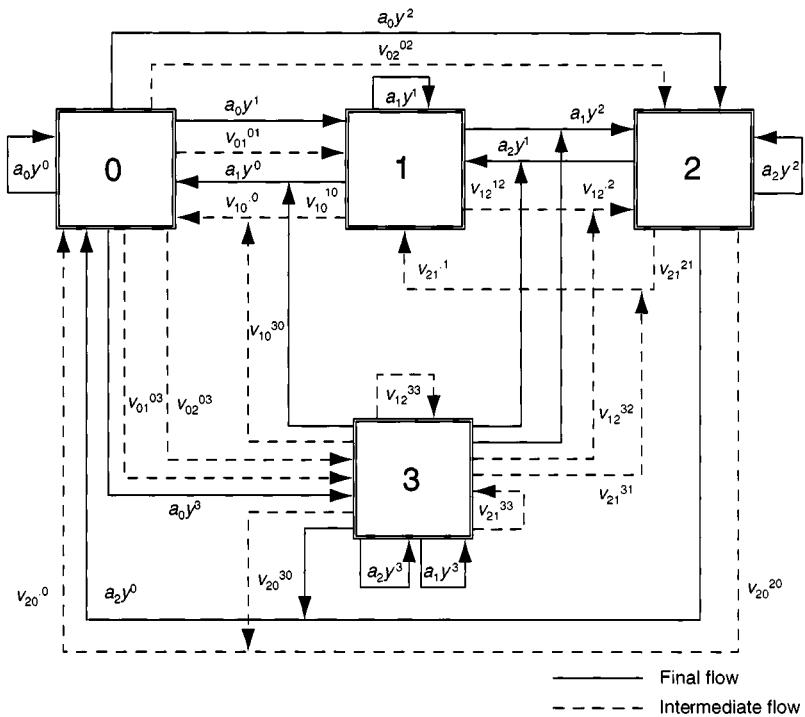


Figure 3. Spatial delivery relations.

The model has been refined by introducing economies of scale and scope, prices, dynamics, and by rendering the number of production units variable; it is as such computable, the underlying reasoning staying the same. The introduction of interpersonal contacts, however, points to issues addressed in the next section.

As an empirical illustration, some results obtained by Stough et al. (2000) should be mentioned here; they computed the distances between the gravity centers of high-tech activities in the Austin (A), Boston (B) and Washington, DC (W), standard metropolitan statistical areas, and used aggregated summed input coefficients (SICs); Table 1 shows the aggregation.

The simple squared correlation coefficients between distances and the SICs were  $-0.5008$  (A),  $-0.7335$  (B), and  $-0.3890$  (W), showing all of them as having negative relationships between distances and SICs, as expected. The strongest relation was observed in the Boston area, the one having the longest industrial history of the three areas investigated, allowing a more appropriate long-term adaptation. This should only be considered as a first intermediate study, later research being conducted at a much more disaggregated level.

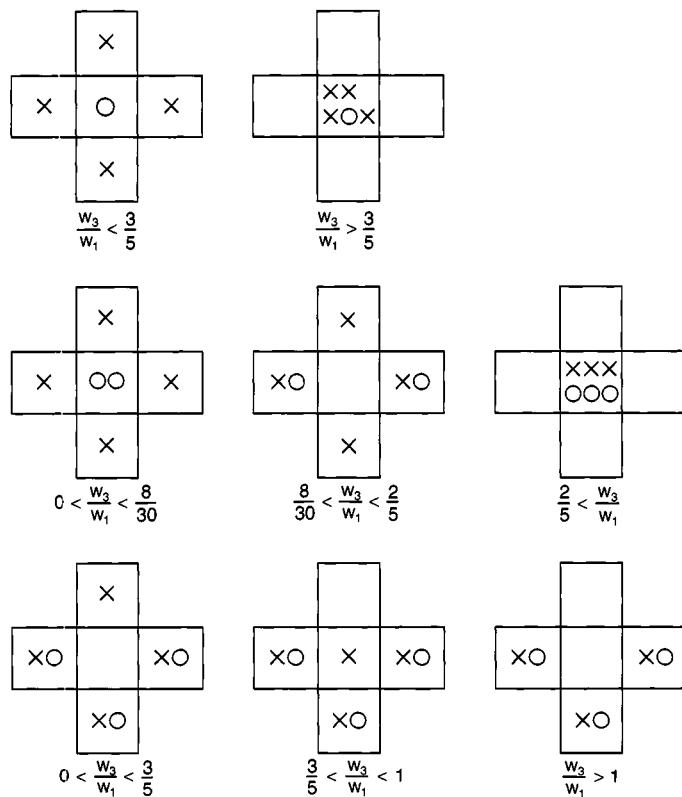


Figure 4. Different economic landscapes on a Manhattan grid of radius 1.

#### 4. Non-market approach

The analysis of Section 3 is in fact limited to market workings in terms of quantities and prices, and even if communication and/or transaction costs are included, they have to be added to the final market values of goods and services.

Externalities, however, add a new element to the approach; they concern all relevant elements that accrue to firms without being traded in markets and are freely accessible to those firms; examples are the local presence of a plentiful, qualitatively appropriate, labor market, of complementary demands, of diffuse information, and of specialized services.

But recently (Steiner, 1998; Green and McNaughton, 2000), direct inter-firm links have been added; this is mainly the result of the integration of spatial and industrial economics, and concern what is called "organizational proximity," in which face-to-face contacts play an essential role, in particular when it comes to

Table 1  
Aggregation of sectors

Input-output code	Sector	SIC
51	Computer and office equipment	357
56	Audio, video, and communication equipment	365–366
57	Electrical components and accessories	367
66	Communications, except radio and television	481, 482, 484, 489
73A	Computer and data-processing services	737

transferring so-called “tacit knowledge,” considered as a “localized” phenomenon. One example is the necessity of personal contacts between individuals working in industry and those in university research facilities.

This redirected approach has led to a large variety of cases. From the point of view of the contents of local concentrations of activities, one finds knowledge clusters (the term “cluster” is systematically used here), vertically integrated clusters, technology clusters, sectoral clusters, and complementary demand clusters; and from the functional point of view one encounters industrial districts, new industrial spaces, innovative milieus, and regional innovative systems.

Those approaches, however, raise a number of questions. The first one is whether transport costs play a more limited role in (joint) location analysis than they did previously; an affirmation often encountered. This is all but certain (Varri Auctores, 1985). To quote from the latter work: concerning road transportation, “Foreseeable gains in the productivity of operations will not necessarily filter through to reduce the cost per unit carried”; for railway hauling, “Given the changes in demand, it is reasonable to assume that the railways will be offering better quality services in the year 2000 at more or less the same cost in real terms”; for inland navigation, “Under moderate foreseeable expectations the cost level in the inland waterway industry might stabilize. The level of freight rates, especially in the cargo operations, will increase.”

Transportation efficiency should be considered, particularly in a globalizing world economy, where new technological systems are made up of lean production and just-in-time deliveries; moreover, it should not be forgotten that the impact on total operating costs is a function of the locational behavior of firms, and a study of the evolution of relative transport and communication costs (the  $a_{ki}$ s of Section 3) is highly advocated. Also, it should be stressed that the new dimensions imply the efficiency of still another facet of “transport,” to wit, the one addressing itself to personal mobility and communication; from that point of view, all improvements of the corresponding systems (intelligent transport systems, fast transit systems), especially inside metropolitan areas where clustering is essentially developing itself, will add to the relative attractiveness of some of

those areas compared with the others. How complicated the matter becomes is illustrated by the possible combinations of different channels: personal contacts with conferences, teleconferencing with formal meetings, etc.

Externalities have always been considered an important locational factor, and so they remain with two important additions: organizational – necessarily interpersonal – contacts, and innovative milieus or regional systems.

As to the first – horizontal integration – some instances have long been known to be important. For example, the “30 miles radius” in the London area in the UK guarantees, especially to small and medium-sized businesses, essential mutual contacts and access to service industries. New, however, are the specific instances derived from access to high-tech information and communication activities (electronic data interchange, new technologies of information and communication), hinted at above.

As to innovative milieus – note that they should be considered as externalities – they are in fact jointly realized by the firms present in a cluster, where they are themselves responsible of their environment. The latter can be fostered by public authorities, which adds an institutional dimension to clustering. But even if some industries might have become “footloose” in the classical market sense, they are not necessarily location-independent from the point of view of the two other dimensions quoted in this section. This leads to the necessity of jointly modeling all the relevant location factors, an example of which is presented in the next section.

## 5. Synthesis

In fact, it can be said that the above developments lead back to a general theory of joint plant location. Figure 3 can then be generalized to Figure 5, where the following elements have been introduced:

- externalities (E, complementary demands next to usual final demand F; labour markets, tacit information, etc.);
- interactivity links of different nature complementing input–output relations (technology linkages, information and knowledge channels, etc.).

On that basis a condensed locational choice model can be built around three factors: transport efficiency, presence of potential inter-firm contacts, and externalities. A remarkable feature of spatial economics to be added is the limiting workings of explanatory factors. In other words, location factors should be present at certain minimal levels for a site to be selected as a potential location. A numerical example will illustrate this (generalization to non-numerical multi-criteria analysis, an important technique in location analysis, is also possible). Suppose a management preparing a locational decision aims at maximizing a

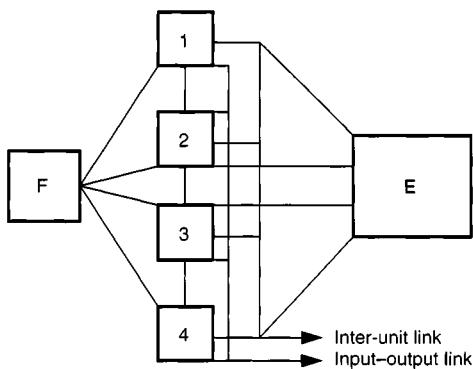


Figure 5. Generalized C graph.

Table 2  
Locational data

<i>w</i>	<i>r</i>	<i>s</i> <sub>1</sub>	<i>s</i> <sub>2</sub>	<i>s</i> <sub>3</sub>	<i>s</i> <sub>4</sub>
4	3	5	8	2	4
8	4	3	6	6	5
6	5	9	4	8	6

weighted sum of location factors (its “objective function”), with a constraining threshold on the presence of every one of them. Then, Table 2 presents the weights (*w*), the minimal requirements (*r*), and the availability of other factors in four relevant sites (*s<sub>i</sub>*, *i* = 1, 2, 3, 4). All figures being chosen are on a 0–10 scale.

The solution is that given by the following mathematical program:

$$\max_x \varphi = w \times Sx \quad (4)$$

s.t.

$$Sx \geq r, \quad (5)$$

$$i \times x = 1, \quad (6)$$

$$x = \hat{x}x. \quad (7)$$

The vectors *w* and *r* correspond to those of Table 2; *x* (with  $\hat{x}$  the derived diagonal matrix) is a vector of decision variables with binary (0–1) values (condition (7)) and *S* is the matrix made up from the four *s<sub>i</sub>* vectors. Expression (4) is the objective function, and condition (5) corresponds to the constraint that minimal levels of the three location elements should be present. Condition (6) expresses the fact that only one site should be selected.

The mathematical program presented is one case of a binary program that can be easily solved. Indeed, the respective values of the objective function are 98, 104, 104, and 92. However site 4, with the lowest value, will be chosen, as all the other sites violate one of the minimal requirements (constraints (5)).

The above reasoning has important implications for regional development policy, especially when less well-equipped regions are envisaged as objectives of that policy. The main difficulty is the absence of one or more of the strategic location factors, which means that traditional stimulation measures (fiscal facilities, cheap loans, etc.) will reveal themselves largely ineffective. Of central importance is the knowledge of the gap between the locally available factors and the minimal requirement constraints of potential investors. From a policy perspective this is a gap that should be filled immediately.

## 6. Conclusions

A new typology finally emerges from the foregoing analyses. Figure 6 presents it by classifying the various types of clusters mentioned in Section 4.

Two new notions are introduced for clarification: clans, which transpose definition 2 of Section 2 to "contact coefficients"  $c_{kl}$  and clubs, which generalize the idea of corps to the same coefficients.

Figure 6 also indicates that mixed local concentrations can occur (see the two-way arrows); a research program for the future includes the identification of the relative strength of the locational forces at work, following the principles of spatial econometrics. Such an analysis is indispensable to find efficient (multi)regional policies. As an example of its feasibility the following model is presented.

Define:

- $\mathbf{V}$  – an  $rxn$  ( $r$  regions,  $n$  sectors) matrix of value added estimates, divided by its sum total (to ensure concavity of expression (8) below);
- $\mathbf{B}$  – an  $nxm$  matrix of binary factors, combining sectors that obey common (implicit or revealed) locational factors;
- $\mathbf{F}$  – an  $rxm$  matrix of those locational factors (elements  $f_{rm}$ ).

The model aims at estimating the elements of  $\mathbf{F}$ , elements that will be given an interpretation afterwards. The model is specified as follows:

$$\max_{f_{rk}} - \sum_{r,k} f_{rk} \ln f_{rk} \quad (8)$$

s.t.

$$\mathbf{VB} = \mathbf{F}, \quad (9)$$

$$\mathbf{Bi} = i. \quad (10)$$

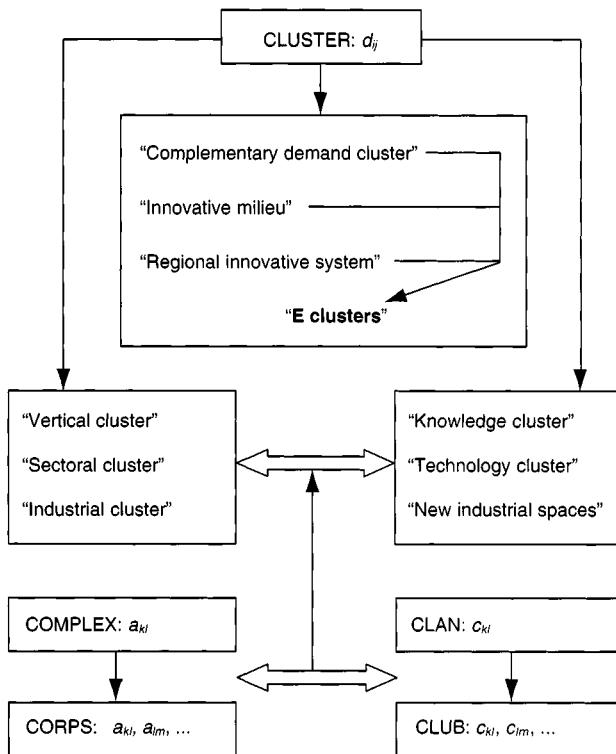


Figure 6. Conceptual flow diagram for 5C analysis.

Function (8) is an entropy function, evaluating the information contained in the sample, which should be maximized, and the  $\mathbf{is}$  represent unit column vectors.

As activity structures in neighboring regions affect the locational advantages of given regions, equation (9) should be generalized to

$$\left[ \left( \mathbf{I} + \sum_j \rho_j \mathbf{C}_j \right) \mathbf{V} \right] \mathbf{B} = \mathbf{F}, \quad (11)$$

where the  $\mathbf{C}_j$  are the  $j$ th degree  $n \times r$  contiguity matrices (showing their degree of neighborhood) of the system of the regions studied, and the  $\rho_j$ s are a decreasing series of coefficients defined on an open 0–1 interval.

The simple model (8) through (10) has been applied to a 1993  $\mathbf{V}$  matrix, derived from data retrieved from the StatLine system of the Dutch Statistical Office.

Tables 3 and 4 reproduce the results for  $\mathbf{B}$  and  $\mathbf{F}$  ( $\times 1000$ ) for, respectively, 12 Dutch provinces and 12 sectors involved in the analysis.

Table 3  
Matrix **B**

Sectors ↓	<i>b</i> <sub>1</sub>	<i>b</i> <sub>2</sub>	<i>b</i> <sub>3</sub>
Agriculture, etc.	0	1	0
Oil, gas	0	1	0
Industry	1	0	0
Building	0	0	1
Trade	0	1	0
Transport, communications	0	1	0
Banking, insurance	0	1	0
Production services	1	0	0
Public authorities	0	0	1
Health care	0	0	1
Culture, recreation	0	0	1
Other services	0	0	1

Table 4  
Matrix **F**

Province ↓	<i>f</i> <sub>1</sub>	<i>f</i> <sub>2</sub>	<i>f</i> <sub>3</sub>
Groningen	12.08	23.36	10.59
Friesland	11.06	12.11	9.44
Drenthe	8.50	8.80	7.09
Overijssel	24.71	17.15	17.59
Flevoland	3.83	4.38	3.29
Gelderland	4.19	3.10	3.34
Utrecht	26.63	25.42	22.56
Noord-Holland	62.36	67.75	47.43
Zuid-Holland	93.49	75.74	64.55
Zeeland	10.49	6.65	5.87
Noord-Brabant	68.04	41.39	37.14
Limburg	27.94	17.92	18.00

The first factor is obviously related to industry and associated services, so it could be considered an “externality factor” from Section 5. The second factor is in fact a combined construct, grouping natural and service elements, and should be partitioned. The third factor relates to final services, and pictures the attraction of large population centers. The provinces Noord-Holland and Zuid-Holland, in the so-called “Rimcity,” are the economic core of the country, and have large factor weights that they partly share with Noord-Brabant. Though at this aggregate level – the sectors analyzed differ from the small sectors to be considered in 5C analysis (see Figure 6) – the picture might be blurred, the technique has been

shown to be operational and already reveals some characteristic traits of the Dutch spatial economy. To appreciate the 5C phenomenon, it will necessarily have to be applied to larger data sets, which is, of course, one of the challenges of future economic cluster analysis.

## References

- Czamanski, S. (1974) *Study of clusters of industries*. Halifax: Institute of Public Affairs, Dalhousie University.
- Green, M.B. and R.B. McNaughton, eds (2000) *Industrial networks and proximity*. Aldershot: Ashgate.
- Paelinck, J.H.P. (2001) "Tinbergen-Bos systems: a compendium of recent research," Working paper (available from the author).
- Paelinck, J.H.P. (2004) "Experiences with input-output and isomorphic analytical tools," in: E. Dietzenbacher and M. Lahr, eds, *Wassily Leontief and input-output analysis*. Cambridge: Cambridge University Press.
- Steiner, M., ed. (1998) *Clusters and regional specialisation. On geography, technology and networks*. London: Pion.
- Stough, R.R., J.H.P. Paelinck and R. Kulkarni (2000) "Industrial clustering: spatial analysis," in: *Annual NARSA Conference*, Paper. Chicago.
- Stough, R.R., J.H.P. Paelinck, R. Kulkarni and G. Yang, (2002) "Activity cluster analysis revisited: theoretical prolegomena and preliminary results," in: *Annual PRSA Conference*, Paper. Bali.
- Varii Auctores (1985) *Foreseeable cost trends in freight transport*. Paris: European Conference of Ministers of Transport.

*Chapter 8*

## CONNECTING MASS TRANSIT AND EMPLOYMENT

THOMAS W. SANCHEZ

*Virginia Polytechnic Institute and State University, Alexandria, VA*

### 1. Introduction

The purpose of this chapter is to highlight the dynamics of employment-related travel. The focus is placed on how transit can connect workers with their places of employment, while acknowledging a broadened definition of “commuting” that accounts for transportation mobility needs for the entire workday rather than simply the trips to and from the workplace. Implicit to this perspective is the consideration of trip purposes beyond those at the beginning and end of the workday. Workers often have complex travel patterns that include travel throughout the day for personal business that are not generally addressed by typical employment accessibility efforts. For this reason, this chapter also discusses employment- and non-employment-related travel for planning and evaluation efforts related to public transit.

According to the 2001 *National Household Travel Survey* (US Department of Transportation, 2001), approximately 33% of all daily travel by persons aged 16–65 years in the USA is work related.<sup>a</sup> Of these trips, about 10% are home based with the remaining trips originating from elsewhere. In 1995, the average US work trip was 1.6 km in length lasting 18.0 min compared with non-work trips, which averaged 11.3 km and 13.4 min. In 2001, the average work trip was 19.2 km and 22.8 min, while the average non-work trip was 11.6 km and 16.5 min. Work trips by transit in 2001 are longer both in terms of distance and time (17.5 km and 49.5 min) compared with non-transit work trips (13.0 km and 21.7 min). Only 2.8% of work trips in the USA are by transit; however, this includes all commuters, whether public transit is available to them or not.<sup>b</sup> The small proportion of transit commuters is likely a better indication of the lack of public transit service in the USA rather than the lack of demand.

<sup>a</sup>For the purposes of this chapter, work-related travel includes all trip activities with the workplace as the destination or the origin.

<sup>b</sup>This includes only travel to the workplace.

Public transit provides increased personal mobility along with local and regional accessibility. It meets the travel needs of persons who have no other means of travel (i.e. they do not own an automobile), and is used as a substitute for the automobile when it is perceived as being more economical or more convenient. Long travel distances, traffic congestion, and high parking costs are disincentives for automobile users that increase the utility of public transit. Public transit is generally provided in medium and large urban and suburban locations, but considerably less so in exurban or rural areas. Land use patterns, particularly residential and employment densities in suburban, exurban, and rural areas, are difficult to serve with fixed-route transit because of the geographic dispersion of trip origins and destinations. Given the pervasiveness of low-density development anticipated in the USA, it is estimated that only a small percentage will occur at densities conducive to transit usage there (Nelson, 2004).

## 2. Elements of travel demand

There are several factors that influence travel activities. Employment-related trips have specific characteristics that distinguish them from other trip purposes. These primary trip-making characteristics were outlined by Meyer and Miller (2001): The following is a brief discussion of each of these in the context of employment-related travel.

- trip purpose;
- trip timing;
- trip origin;
- trip destination;
- available modes;
- available routes;
- trip frequency.

Each of these factors does not independently influence trip-making behavior; rather, there are several interdependencies and complex relationships among land use and traveler characteristics. This discussion is not intended to explain the traditional travel-demand modeling process but rather to illustrate the association between each element and employment-related travel, as mentioned in the introduction.

### 2.1. *Trip purpose*

Work trips (especially in the direction of the workplace) typically require a reliable schedule for departure and arrival. Having employees arrive at work on

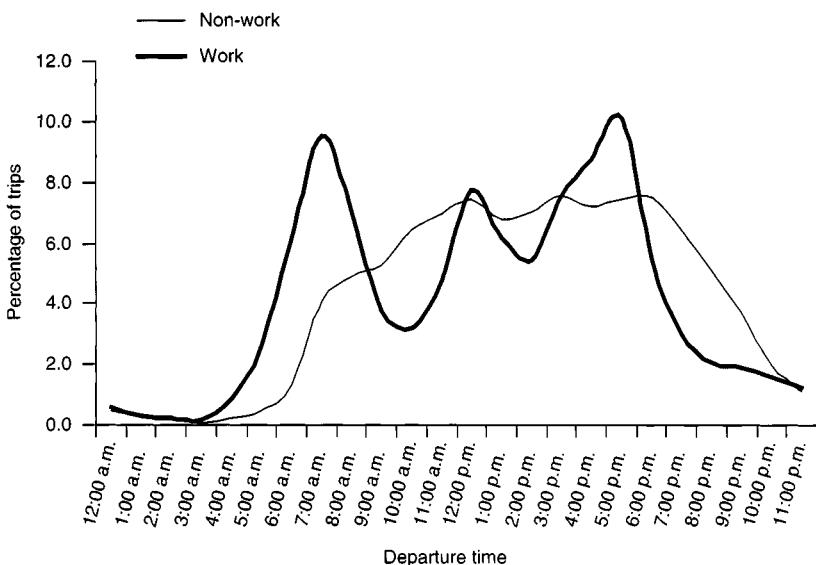


Figure 1. Trip time distribution in the USA. (Source: US Department of Transportation, 2001.)

time is important to businesses and employers. However, the return trip home may not have the same temporal rigidness, unless the commuter has other personal or household requirements – such as picking up a child at day care or school. The return trip may need to have a flexible schedule in the event that a worker needs to leave work early or late, often a function of job-related factors that vary work day completion.

Besides trip segments specifically for reaching or returning from the workplace, other trip purposes during the workday include travel for meals, errands, and medical appointments. These trip purposes can be associated with trip chains before work hours, after work hours, or during the workday.

## 2.2. Trip timing

As mentioned, the trip purpose is a major determinant of the temporal pattern of departure times (as well as desired arrival times). This is best illustrated by the trimodal distribution of trip making associated with work-related travel activities (Figure 1). The three peaks in travel activity occur in relation to the journey to work, midday travel for lunch or errands, and the journey from work. Non-work-related trips tend to occur later in the day and during the evening hours, compared

with work trips. It is also true that a significant proportion of work-related travel (53.1% from Figure 1) occurs outside of peak times, when transit service frequencies tend to be lower (US Department of Transportation, 2001). This has negative impacts on workers with second- or third-shift work schedules – making public transit an unreliable source of mobility for them.<sup>a</sup>

The reasons for workplace-related trips undertaken between 11:00 a.m. and 2:00 p.m. in Figure 1 are divided primarily among shopping (17.6%), seeing family members and other persons (11.6%), going for meals (10.0%), and social/recreation (9.3%). It is important to recognize that these trip types vary from those of the morning and afternoon peak periods, and thus have different travel requirements in terms of trip duration, trip distances, and cost. It is likely that a significant proportion of workers drive to work simply because they have no other means to undertake these ancillary trips, while the home-to-work and work-to-home routes are adequately served by transit.

### *2.3. Trip origins and destinations*

The spatial distribution of trip origins and destinations is another fundamental element of transit route planning. This means that areas with high densities of trip origins or destinations typically have the highest demand for transit services. In addition to exhibiting high levels of demand, dense residential or employment zones are also the most easily served because greater numbers of potential riders and locations can be served per operating unit cost compared with low-density areas. The continuum of density levels is illustrated by the differences between rural zones (low) and the urban core (high) (Figure 2).

### *2.4. Trip mode*

Just as the trip purpose influences trip schedules, origins, and destinations, so it also influences the mode of travel. Mode choice is obviously affected by the set of modes available to a traveler. If the traveler does not own or have access to an automobile, it is more likely that trips will be made by transit – or in some cases not made at all. If a traveler does not live within a reasonable distance of a transit stop, it is less likely that trips will be made by transit. The level of access to different transportation modes defines personal mobility. The continuum of personal transportation mobility can be characterized at the lowest level as walking and at

<sup>a</sup>This accounts for approximately 20% of work trips occurring near the afternoon peak, with return trips in the early hours prior to the morning peak.

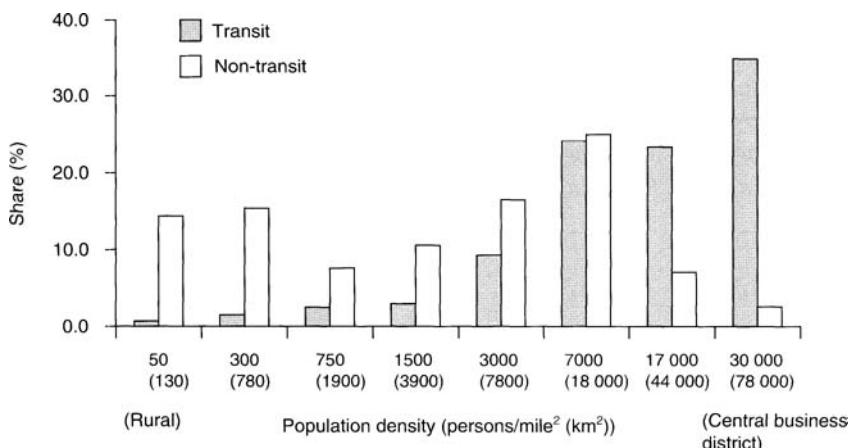


Figure 2. Population density (at the origin) and transit usage in the USA. (Source: US Department of Transportation, 1995; note that census data for population density were not available in the first release of the *National Household Travel Survey*.)

the highest level as automobile ownership (especially personally owned autos with unlimited availability). The range of personal mobility greatly influences when, where, and how a person will travel. Other factors also include the ability of a person to use these modes, which are affected by their age, health, and licensure status.

There are also several other considerations related to the effectiveness and efficiency of certain modal options for employment-related travel. Transit modes vary in the advantages that they represent in terms of passenger capacity, travel speeds, operation costs, and environmental impacts (Black, 1995). For example, rail transit facilities have a higher potential to attract complementary land uses, offer higher levels of comfort, and greater vehicle capacities. On the other hand, bus transit provides more route and service flexibility, and relatively low capital costs compared with rail. A mix of service types appropriate to local and regional characteristics can prove more effective than a single-service-type approach.

## 2.5. Available routes

Vehicular travel cannot occur unless there is appropriate infrastructure. A road network comprises a hierarchy of facilities that accommodate a range of travel speeds and volumes. Travel typically occurs along a route that represents the lowest cost in terms of speed, convenience, and expense when high levels of demand exist. Competition for space on the network also produces congestion or

high travel costs. Again, depending on the trip purpose, travelers balance the cost associated with particular mode choices along with route selection (which assumes that choices are available). Subsidies and incentives can be used to affect capacity (i.e. reduce impedance) by shifting departure times, travel modes, and routes selected. Again, these are most effective when a traveler has the flexibility to make alternative trip choices.

## 2.6. *Trip frequency*

The issue of trip frequency, especially for individuals, is another facet of travel-related decision-making. Trips made on a weekly, monthly, or irregular basis (such as shopping, recreation, and medical appointments) may afford a wider range of travel options compared with trips that are made on a rigid daily basis (such as the work commute). Work trips tend to involve less flexibility in scheduling and demand higher levels of reliability compared with shopping or leisure trips. These trips have more flexibility for departure and return times – less so in the case of work trips for reasons already mentioned. Like other elements of travel demand, the frequency of trip making at the individual and aggregate level is an important indicator of the types of transportation service needed to satisfy local and regional travel demands.

Figure 3 deviates from the standard four-step model that typically depicts an iterative, linear process. The diagram conveys the simultaneity of factors and how they affect decision-making and travel behavior not explicit in the four-step model.

## 3. Work trip factors

The preceding discussion about the factors affecting trip making provides a general foundation that varies depending on trip purpose. As mentioned, work commute trips place relatively more weight on timeliness and reliability, and therefore can be viewed with a more specific set of criteria. In planning employment-serving transit the following factors (derived from the previous discussion) are primary concerns:

- the distribution of departure and return trip times, and also the day of the week (trip time);
- the direction of trip flows (destination routes);
- the modal availability and intermodal connections (available modes).

Each factor is briefly discussed below as it pertains to employment-related travel.

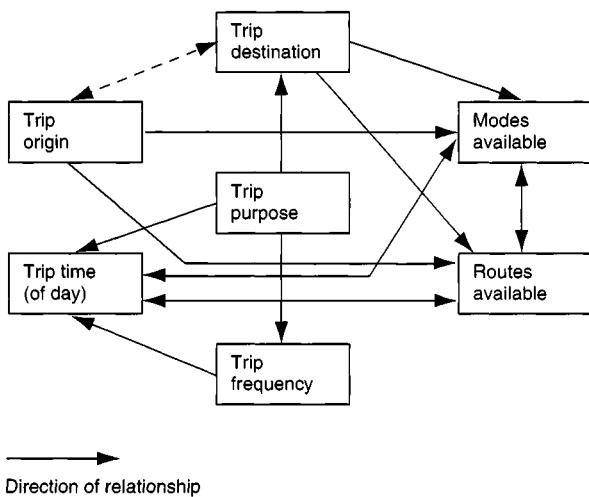


Figure 3. Diagram summarizing work travel decision-making.

### 3.1. Distribution of trip times and day of week

Work trips most commonly occur in relation to the standard business schedule – with peaks occurring in the morning (6–8 a.m.) and evening (4–6 p.m.). This varies, however, depending on the occupations of workers in a particular service area. White collar and professional workers tend to adhere to the traditional, trimodal peak hour schedule, while blue collar and service workers have a higher likelihood of second, third, or weekend shift hours. US data in Figures 4 and 5 show that lower-skilled workers (those with lower education levels) have a higher likelihood of working earlier and later shifts compared with higher-skilled workers (those with higher education levels). In addition, workers with less than a high-school degree were twice as likely to make work trips on weekends compared with those with graduate or professional school degrees. Varied work schedules and commute travel times also occur between full- and part-time employment by urban location. For this reason, occupations, which are correlated with socio-economic status, are an important determinant of transit demand for employment-related travel.

### 3.2. Direction of trip flows

As mentioned, the occupational characteristics of a service area affect the demand for when transit demand will occur. The occupational characteristics of residents

also affect the destinations for transit service. Employment accessibility measures that account for differences in employment types (by industrial classification) as well as turnover rates are useful in predicting the travel direction of commute flows (Shen, 2001). In essence, these measures indicate urban location relative to

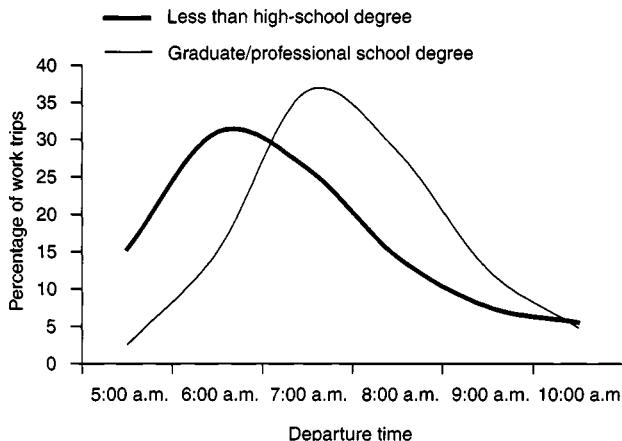


Figure 4. Correlation between skill level (education) and work travel time (morning peak) in the USA. (Source: US Department of Transportation, 2001.)

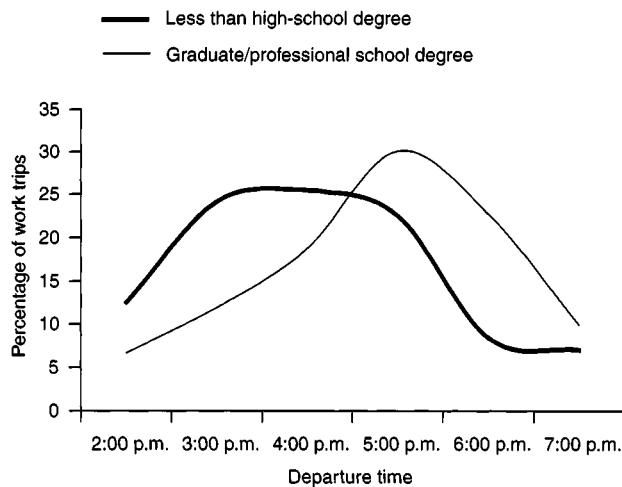


Figure 5. Correlation between skill level (education) and work travel time (afternoon peak) in the USA. (Source: US Department of Transportation, 2001.)

**Table 1**  
Density recommendations for varying levels of service in the USA

Service	Frequency	Coverage	Dwelling units/ acre (ha)
Rapid transit (rail)	5 min peak headway	100–150 mile (260–390 km <sup>2</sup> ) corridor	12 (5)
Light rail	5 min peak headway	25–100 mile (64–260 km <sup>2</sup> ) corridor	9 (4)
Bus – frequent service	120 buses/day	0.5 mile (0.8 km) between routes	15 (6)
Bus – intermediate service	40 buses/day	0.5 mile (0.8 km) between routes	7 (3)
Bus – minimum service	20 buses/day	0.5 mile (0.8 km) between routes	4 (2)

*Source:* Pushkarev et al. (1982).

downtowns or other employment concentrations, and serve as general indicators of potential travel interaction.

### 3.3. Modal availability

The type of transit services provided (bus, subway, light rail, etc.) also influences potential commute characteristics. Land use patterns and associated densities influence the feasibility of commuter rail, light rail, and subway – especially when pedestrian access is considered. Pedestrian (for bus and rail) and vehicular access (including parking for rail) design factors play important roles in mode choice decisions. A distance of 0.5 km, depending on weather conditions and physical geography, is usually considered the maximum likely distance that people will walk to use transit. These distances vary depending upon the mode and facilities serving each (see Table 1 for recommended service types). This is especially true when parking is available at transit stops – where people may drive rather than walk to use the system. Along with the proximity to transit stops, the number of connections needed to reach a destination should be minimized. Multiple transfers decrease the convenience of transit, with more than two transfers having a negative effect on riders' willingness to travel by transit.

Figure 6 summarizes the decision-making elements of work commutes. The primary consideration is the level of employment access from the trip origin. High levels of access decrease the potential travel time and cost by transit while low levels increase travel time and cost. Employment access is particularly affected by the availability of transit, but the provision of transit service is also influenced by

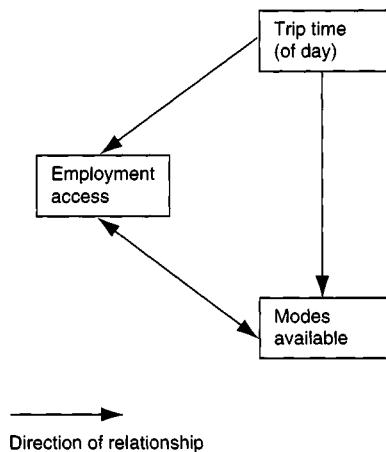


Figure 6. Work commute diagram.

the proximity of employment or other destinations. The time and day of travel dictates employment accessibility levels where the travel time and cost may be less during non-peak periods provided that service is available.

#### 4. Other factors

##### 4.1. Network characteristics

Underlying employment accessibility and levels of transit availability are transit network connectivity and extensiveness. As network connectivity increases, the likelihood that two locations are accessible by transit increases. A well-connected network, however, may not adequately serve the geographic extent of an urban or metropolitan area. This is especially true where significant levels of suburban employment need to be accessed by workers from the urban core, as is the case for reverse commuters. In the case of suburb-to-suburb commuters, effective transit networks must be both extensive and well connected to match the relative dispersion of residential and employment concentrations.

##### 4.2. Network extensiveness

Because a majority of large transit systems emanate from the central business district, suburban service levels tend to be relatively low. With lower residential

and employment densities, transit routes increase in extent in order to achieve service frequencies comparable with urban core areas. Large metropolitan areas tend to have multiple transit systems to serve suburban communities and regional commuting sheds. Coordination among these transit service providers is necessary for effective and efficient service provision. National and regional rail, bus, and subway systems in western Europe are an example of extended networks based on such intermodal connections.

#### *4.3. Network connectivity*

Well-connected transit systems exist when supported by appropriate land use patterns. Dense, flat, grid-based areas that developed in older US cities were easier to connect with street car systems that relied on overhead electrical lines that followed urban street patterns. As the street car was replaced by bus transit, the potential for route flexibility increased as well as system extensiveness. However, as systems extend further into the suburbs, connectivity levels decline due to increased geographic coverage and route dispersion. Early streetcar systems were successful because they served more compact development patterns, even for lines that extended further out into the urban fringe. Development patterns, especially higher densities, were major factors influencing the success of these systems.

#### *4.4. Physical access – walking distances*

At the urban scale, the extensiveness and connectivity of transit systems are related to pedestrian access levels to stations or stops. As mentioned earlier, walking distances over 0.5 km discourage pedestrian access to transit. Little research, however, has focused on how increased distance to transit stops (i.e. transit service delivery) effect labor participation and economic opportunity. Sanchez (1999) analyzed the relationship between transit access and employment levels for the US cities of Portland, Oregon, and Atlanta, Georgia. He found a statistically significant and positive relationship between physical proximity and the average annual number of weeks worked by persons between the ages of 16 and 65 years (Figure 7).<sup>a</sup> The results suggest that proximity to bus transit was more important compared with light rail access in both cities. This appears reasonable

<sup>a</sup>Controlling for employment accessibility, vehicle ownership, educational attainment, neighborhood racial composition, household composition, occupations, and time leaving for work.

given that the light rail systems in Portland and Atlanta are not extensive and have low levels of connectivity compared with their respective bus systems.

#### 4.5. Vehicle ownership levels

Another important factor affecting both transit usage, mobility, and employment accessibility are rates of vehicle ownership. Obviously, households that own more autos are more mobile, which translates into higher levels of social and economic opportunity. Automobiles generally provide superior mobility and accessibility compared with fixed-route urban transit systems. Public transit can, however, be quite successful when linking high-density origins and destinations, especially with significant traffic congestion and high parking costs. Figure 8 indicates the level of transit usage for work trips by individuals in households that do not own cars. Overall, the likelihood of using transit declines significantly from over 30% to under 10% when vehicles are owned by household members. This pattern varies based upon metropolitan area population size as shown.

### 5. Summary

The previous discussion identifies several important factors that should be considered when planning or evaluating employment-serving transit. Besides the

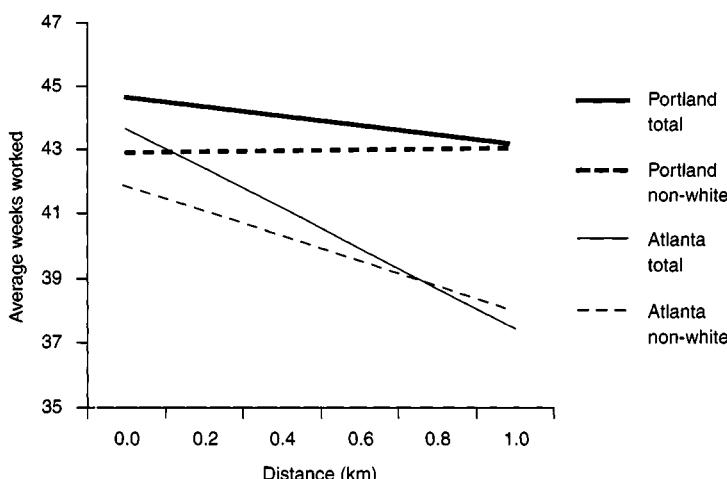


Figure 7. Walking distance to the nearest transit stop and employment levels for Portland and Atlanta. Note: only census block groups within city limits and within a mile (1.6 km) of a transit service are included.

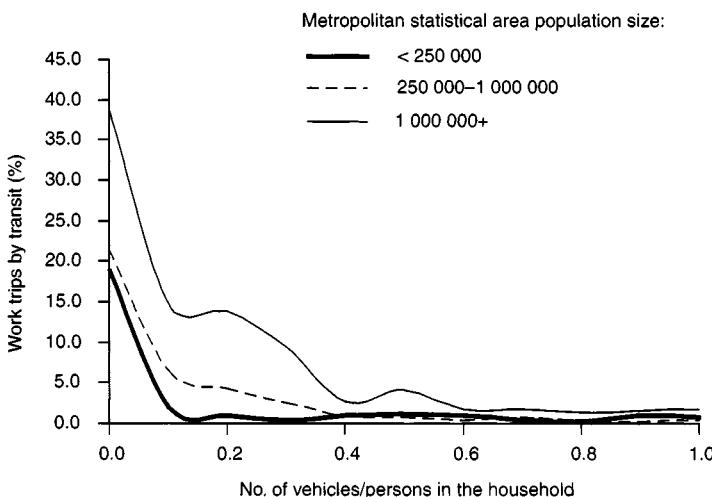


Figure 8. Transit commuting and vehicle ownership in the USA. (Source: US Department of Transportation, 1995.)

traditional considerations such as trip purpose, time of trip, trip origin, etc., it is useful to consider the geographic coverage and connectivity of a transit system in relation to linking residences with employment concentrations. Worker travel needs extend beyond the trips to and from the job site, and should include overall mobility requirements to make transit more attractive and beneficial. In addition, physical access to transit stops and vehicle ownership rates are important micro-level factors that affect system performance. Employment-serving transit can be more efficient, effective, and equitable if it addresses a range of community and regional accessibility and mobility needs.

## References

- Black, A. (1995) *Urban mass transportation planning*. New York: McGraw Hill.
- Meyer, M.D. and E.J. Miller (2001) *Urban transportation planning: a decision oriented approach*, 2nd edn. New York: McGraw Hill.
- Nelson, A.C. (2004) *Development needs of the next generation*. Washington, DC: National Association of Realtors.
- Pushkarev, B.S., J.M. Zupan and R.S. Cumella (1982) *Urban rail in America: an exploration of criteria for fixed-guideway transit*. Bloomington: Indiana University Press.
- Sanchez, T.W. (1999) "The connection between public transit and employment," *Journal of the American Planning Association*, 65:284–296.
- Shen, Q. (2001) "A spatial analysis of job openings and access in a US metropolitan area," *Journal of the American Planning Association*, 67:53–68.

US Department of Transportation (1995) *1995 national personal transportation survey*. Washington, DC: Federal Highway Administration.

US Department of Transportation (2001) *2001 national household travel survey*. Washington, DC: Federal Highway Administration.

***Part 3***

**LAND USE AND TRANSPORTATION MODELING**

## OVERVIEW OF LAND USE TRANSPORT MODELS

MICHAEL WEGENER

*Urban Regional Research, Dortmund*

### 1. Introduction

The previous chapters in this handbook have shown that spatial development, or land use, determines the need for spatial interaction, or transport, but that transport, by the accessibility it provides, also determines spatial development. However, it is difficult to empirically isolate impacts of land use on transport and vice versa because of the multitude of concurrent changes of other factors. This poses a problem if the likely impacts of integrated land use and transport policies to reduce the demand for travel are to be predicted.

There are principally three methods to predict those impacts. The first is to ask people how they would change their location and mobility behavior if certain factors, such as land use regulations or transport costs, would change ("stated preference"). The second consists of drawing conclusions from observed decision behavior of people under different conditions on how they would be likely to behave if these factors would change ("revealed preference"). The third method is to simulate human decision behavior in mathematical models.

All three methods have their advantages and disadvantages. Surveys can also reveal subjective factors of location and mobility decisions; however, their respondents can only make conjectures about how they would behave in still unknown situations, and the validity of such conjectures is uncertain. Empirical studies based on observation of behavior produce detailed and reliable results; these, however, are valid only for existing situations and are therefore not suited for the assessment of novel, untested policies. In addition, it is usually not possible to associate the observed changes of behavior unequivocally with specific causes, because in reality several determining factors change at the same time.

Mathematical models of human behavior are also based on empirical surveys or observations. The difference is that the conclusions to be drawn from the survey and observation data are quantified. Strictly speaking, the results of mathematical models are no more universally valid than those of empirical studies but are only valid for situations that are similar to those for which their parameters were

estimated. Nevertheless it is possible to transfer human behavior represented in mathematical models within certain limits to still unknown situations. In addition, mathematical models are the only method by which the effects of individual determining factors can be analyzed by keeping all other factors fixed.

In this chapter recent developments in the field of operational integrated land use transport models will be reviewed, with special emphasis on their ability to test both land use and transport policies and to assess their impacts.

## 2. Existing urban land use transport models

The models reviewed here are integrated, i.e. incorporate the most essential processes of spatial development in urban regions. This implies that they forecast urban land use, where land use denotes a range of uses such as residential, industrial, and commercial. This excludes partial models addressing only one subsystem, such as housing or retail. It is essential that the links from transport to land use are considered; transport itself may be modeled either endogenously or by an exogenous transport model. The models are operational in the sense that they have been implemented, calibrated, and used for policy analysis for at least one metropolitan region.

The number of real-world applications of models falling under the above definition has increased steadily over the last two decades. There has been a continuous reflection of purpose, direction, and theoretical basis of land use transport modeling (e.g. see Mackett, 1985; Berechman and Small, 1988; Boyce, 1988; Harris, 1994; Wegener, 1994, 1998a; Wilson, 1997; Wegener and Fürst, 1999).

To assess the current state of the art in urban modeling, in this section a framework for the classification and evaluation of urban models is first established.

### 2.1. Urban change processes

For the evaluation of operational urban models, the urban change processes to be modeled are identified. Eight types of major urban subsystem are distinguished. They are ordered by the speed by which they change, from slow to fast processes:

- *Very slow change: networks and land use.* Urban transport, communications, and utility networks are the most permanent elements of the physical structure of cities. Large infrastructure projects require a decade or more, and once in place are rarely abandoned. The land use distribution is equally stable; it changes only incrementally.

- *Slow changes: workplaces and housing.* Buildings have a life-span of up to 100 years and take several years from planning to completion. Workplaces (non-residential buildings) such as factories, warehouses, shopping centers or offices, theatres or universities exist much longer than the firms or institutions that occupy them, just as housing exists longer than the households that live in it.
- *Fast change: employment and population.* Firms are established or closed down, expanded or relocated; this creates new jobs or makes workers redundant, and so affects employment. Households are created, grow or decline, and eventually are dissolved, and in each stage in their life cycle adjust their location and motorization to their changing needs; this determines the distribution of population and car ownership.
- *Immediate change: goods transport and travel.* The location of human activities in space gives rise to a demand for spatial interaction in the form of goods transport and travel. These interactions are the most flexible phenomena of spatial urban development; they can adjust in minutes or hours to changes in congestion or fluctuations in demand, though in reality adjustment may be retarded by habits, obligations, or subscriptions.

There is a ninth subsystem: the urban environment. Its temporal behavior is more complex. The direct impacts of human activities, such as transport noise and air pollution, are immediate; other effects such as water or soil contamination build up incrementally over time; and still others such as long-term climate effects are so slow that they are barely observable. All other eight subsystems affect the environment by energy and space consumption, air pollution, and noise emission, whereas only the locational choices of housing investors and households, and firms and workers, are co-determined by environmental quality, or lack of it. All nine subsystems are partly market-driven and partly subject to policy regulation.

In the 1950s, in the USA, the first efforts were made to study the interrelationship between transport and the spatial development of cities systematically. Hansen (1959) demonstrated for Washington, DC, that locations with good accessibility had a higher chance of being developed, and at a higher density, than remote locations. The recognition that trip and location decisions co-determine each other and that, therefore, transport and land use planning needed to be coordinated, quickly spread among US planners, and the “land use transport feedback cycle” became commonplace in the US planning literature. The set of relationships implied by this term can be briefly summarized as follows (see Figure 1):

- The distribution of land uses, such as residential, industrial, or commercial, over the urban area determines the locations of human activities such as living, working, shopping, education, or leisure.

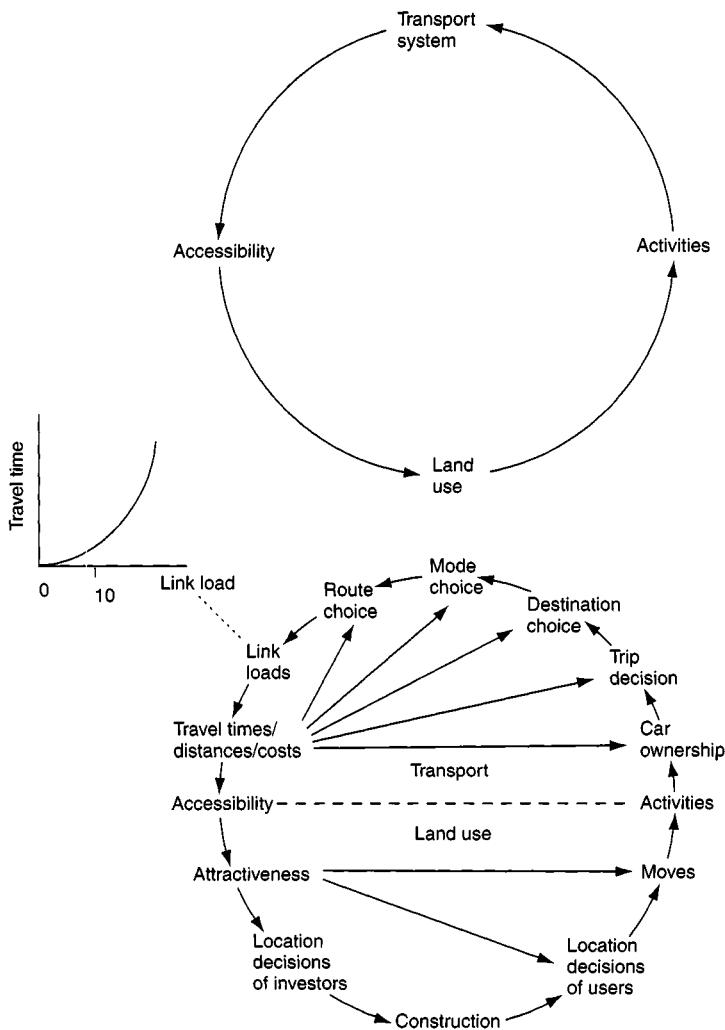


Figure 1. The land use transport feedback cycle.

- 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.
- The distribution of infrastructure in the transport system creates opportunities for spatial interactions, and can be measured as accessibility.
- The distribution of accessibility in space co-determines location decisions and so results in changes of the land use system.

This simple explanation pattern is used in many engineering-based and human-geography-derived urban development theories (see Chapter 4).

Lowry's (1964) *Model of Metropolis* was the first attempt to implement the urban land use transport feedback cycle in an operational model. The Lowry model essentially consists of a residential location model and a service and retail employment location model nested into each other (see Chapter 11). The Lowry model stimulated a large number of increasingly complex modeling approaches, such as the work by Putman (1991). Boyce et al. (1981) developed combined equilibrium models of residential location, mode, and route choice. From these pioneering efforts, a wide range of different approaches to model urban land use and transport have evolved. The following section provides an overview.

## 2.2. Twenty urban models

For this overview, 20 contemporary urban land use transport models were selected for a comparative review; these models represent the current state of the art of urban modeling, though it cannot be excluded that promising new approaches in this rapidly moving field were overlooked:

- BOYCE – the combined models of location and travel choice developed by Boyce (Boyce et al., 1983; Boyce and Mattsson, 1991);
- CUFM – the California urban futures model, developed at the University of California at Berkeley (Landis, 1994; Landis and Zhang, 1998a,b);
- DELTA – the land use/economic modeling package developed by Davids Simmonds Consultancy, Cambridge (Simmonds, 2001);
- ILUTE – the integrated land use, transportation, environment modeling system, under development at several Canadian universities (Miller and Salvini, 2001);
- IMREL – the integrated model of residential and employment location, developed at the Royal Institute of Technology, Stockholm, by Anderstig and Mattsson (1998);
- IRPUD – the model of the Dortmund region developed at the University of Dortmund (Wegener, 1982a,b, 1986b; Wegener et al., 1991);
- ITLUP – the integrated transportation and land use package, developed by Putman (1991, 1998) and consisting of the residential location model DRAM and the employment model EMPAL;
- KIM – the non-linear urban equilibrium model developed at the University of Illinois at Urbana by Kim (1989) and Rho and Kim (1989);
- LIILT – the Leeds integrated land use/transport model, developed at the University of Leeds by Mackett (1991a,b);

- MEPLAN – the integrated modeling package developed by Echenique and others (Echenique et al., 1990; Hunt and Simmonds, 1993);
- MetroSim – the microeconomic land use and transport model developed for the New York Metropolitan Area by Anas (1994, 1998);
- MUSSA – the five-stage land use transport model, developed for Santiago de Chile by Martinez (Martinez, 1992; Martinez and Donoso, 1995);
- PECAS – the production, exchange and consumption allocation system, developed at the University of Calgary (Hunt and Abraham, 2003);
- POLIS – the projective optimization land use information system, developed by Prastacos for the Association of Bay Area Governments (Prastacos, 1986);
- RURBAN – the random-utility urban model, developed by Miyamoto (Miyamoto and Udomsri, 1996);
- STASA – the master-equation-based transport and urban/regional model developed for the metropolitan region of Stuttgart by Haag (1990);
- TLUMIP – the land use transport model of the US state of Oregon developed in the Oregon transport and land use model integration program (Oregon Department of Transportation, 2002);
- TRANUS – the transport and land use model developed by de la Barra (1989, 1998);
- TRESIS – the transportation and environment strategy impact simulator, developed at the University of Sydney by Hensher and Ton (2001);
- UrbanSim – the microeconomic model of location choice of households and firms developed by Waddell (2002).

These 20 models are now compared in terms of the criteria comprehensiveness, model structure, theoretical foundations, modeling techniques, dynamics, data requirements, calibration and validation, operability, and applicability.

### *Comprehensiveness*

All 20 models are comprehensive in the sense that they address at least two of the eight subsystems identified above. Only ILUTE, MEPLAN, STASA, PECAS, TLUMIP, and TRANUS encompass all eight subsystems. IRPUD, LILT, MetroSim, and TRESIS address all subsystems except goods transport, and KIM models goods movements but not physical stock and land use. Half of the models make no distinction between activities (population and employment) and physical stock (housing and workplaces). Six models (DELTA, CUFM, MUSSA, POLIS, RURBAN, and UrbanSim) do not in themselves model transport but rely on interaction with existing transport models. Only DELTA, ILUTE, IRPUD, LILT, and UrbanSim model demographic change and household formation. Table 1 shows the urban subsystems that are modeled with each model.

Table 1  
Urban subsystems represented in land use transport models

Models	Speed of change								
	Very slow		Slow			Fast		Immediate	
	Networks	Land use	Workplaces	Housing	Employment	Population	Goods transport	Travel	
BOYCE	+				+	+		+	
CUFM	(+)	+	+	+	+	+		(+)	
DELTA	(+)	+	+	+	+	+		(+)	
ILUTE	+	+	+	+	+	+	+	+	
IMREL	+	+	+	+	+	+		+	
IRPUD	+	+	+	+	+	+		+	
ITLUP	+	+			+	+		+	
KIM	+				+	+	+	+	
LILT	+	+	+	+	+	+		+	
MEPLAN	+	+	+	+	+	+	+	+	
METROSIM	+	+	+	+	+	+		+	
MUSSA	(+)			+	+	+		(+)	
PECAS	+	+	+	+	+	+	+	+	
POLIS	(+)	+			+	+		(+)	
RURBAN	(+)	+			+	+		(+)	
STASA	+	+	+	+	+	+	+	+	
TLUMIP	+	+	+	+	+	+	+	+	
TRANUS	+	+	+	+	+	+	+	+	
TRESIS	+	+	+	+	+	+		+	
URBANSIM	(+)	+	+	+	+	+		(+)	

Key: (+) provided by linked transport model.

### *Model structure*

With respect to overall model structure, two groups can be distinguished. One group of models searches for a unifying principle for modeling and linking all subsystems; the others see the city as a hierarchical system of interconnected but structurally autonomous subsystems. The resulting model structure is either tightly integrated, "all of one kind," or consists of loosely coupled submodels, each of which has its own independent internal structure. The former type of model is called "unified," the latter "composite" (Wegener et al., 1986). Nine of the 20 models (BOYCE, MUSSA, KIM, MEPLAN, MetroSim, PECAS, RURBAN, TRANUS, and STASA) belong to the unified category, the remaining 11 are composite. The distinction between unified and composite model designs has important implications for the modeling techniques applied and for the dynamic behavior of the models.

### *Theoretical foundations*

In the last 30 years great advances in theories to explain spatial choice behavior and in techniques for calibrating spatial choice models have been made. Today there is a broad consensus about what constitutes a state-of-the-art land use model: Except for one (CUFM), which uses allocation rules, all models rely on random utility or discrete choice theory to explain and forecast the behavior of actors such as investors, households, firms, or travelers. Random utility models predict choices between alternatives as a function of the attributes of the alternatives, subject to stochastic dispersion constraints that take account of unobserved attributes of the alternatives, differences in taste between the decision-makers, or uncertainty or lack of information (Domencich and McFadden, 1975). Anas (1983) showed that the multinomial logit model resulting from random utility maximization is, at equal levels of aggregation, formally equivalent to the entropy-maximizing model proposed by Wilson (1967); he thus laid the foundation for the convergence and general acceptance of formerly separate strands of theory. The STASA model is based on the master equation approach, and may be seen as a dynamic and decision-based multi-agent system (Haag, 1990). Underneath that uniformity, however, there are significant differences between the theoretical foundations of the models:

- Eleven models (DELTA, IMREL, KIM, MEPLAN, MetroSim, MUSSA, PECAS, RURBAN, TLUMIP, TRANUS, and TRESIS) represent the land (or floorspace or housing) market with endogenous prices and market clearing in each period; three (ILUTE, IRPUD, and UrbanSim) have endogenous land and housing prices with delayed price adjustment. These models are indebted to microeconomic theory, in particular to Alonso's (1964) theory of urban land markets or bid-rent theory. The models

without market equilibrium rely on random utility maximization; however, three of the microeconomic models (MUSSA, RURBAN, and STASA) are hybrids between bid-rent and random utility theory. All models with transport submodels use random utility or entropy theory for modeling destination and mode choice, except the STASA model.

- Only KIM and MetroSim determine a general equilibrium of transport and location with endogenous prices. Other models are equilibrium models of transport only (ILUTE, IRPUD, ITLUP, and TLUMIP), of transport and activity location separately (IMREL, MEPLAN, PECAS, TRESIS, and TRANUS), or of transport and location combined but without endogenous prices (BOYCE and LILT). Five models apply concepts of locational surplus (IMREL and POLIS), random utility (DELTA, IRPUD, and ITLUP), or profitability (CUFM) to locate activities. ITLUP may be brought to general equilibrium, but this is not normally done; MetroSim may produce a long-run equilibrium or converge to a steady state in annual increments. STASA describes the short-term redistribution of population during a day due to transport events.
- IMREL uses its equilibrium mechanism to determine the distribution of housing that maximizes locational surplus, and so is a true optimization model, whereas all other models in the sample simulate one particular scenario only. Despite earlier attempts at optimization in urban models (e.g. Brotchie et al., 1980), optimization approaches in urban models have all but disappeared (a recent exception is described in Pfaffenbichler and Shepherd, 2002).
- Several other theoretical elements are built into some models. MEPLAN, MetroSim, PECAS, and TRANUS use export base theory to link population and non-basic employment to exogenous forecasts of export industries. DELTA, ILUTE, IRPUD, LILT, TLUMIP, and UrbanSim apply standard probabilistic concepts of cohort survival analysis in their demographic and household formation submodels. IRPUD also utilizes ideas from time geography, such as time and money budgets, to determine action spaces of travelers in its transport submodel.

### *Modeling techniques*

In all 20 models, the urban region is represented as a set of discrete subareas or zones. Time is typically subdivided into discrete periods of 1 to 5 years. This classifies all models except IMREL (which is static) as recursive simulation models:

- STASA uses a 1 year period for the urban/regional modeling, and a 1 h period for redistribution effects due to transport events. In nine models (BOYCE, IMREL, KIM, LILT, MEPLAN, MetroSim, PECAS, RURBAN,

and TRANUS) transport and location are simultaneously determined in spatial-interaction location models in which activities are located as destinations of trips; in the remaining models (and in the employment location model of IMREL) transport influences location via accessibility indicators. In the models with network representation, state-of-the-art modeling techniques are applied, with network equilibrium the dominant trip assignment method despite its weakness of collapsing to all-or-nothing assignment in the absence of congestion. Only ITLUP, MEPLAN, STASA, and TRANUS have multiple-path assignment allowing for route choice dispersion, and only ILUTE and TLUMIP use activity-based trip generation.

- For representing flows of goods, spatial input-output methods are the standard method. DELTA, KIM, MEPLAN, PECAS, and TRANUS use input-output coefficients or demand functions for intersectoral flows and random utility or entropy models for their spatial distribution. MEPLAN, PECAS, and TRANUS incorporate industries and households as consuming and producing “factors” resulting in goods movements or travel.
- With the exception of CUFM, all models are aggregate at a meso level, i.e. all results are given for medium-sized zones and for aggregates of households and industries. CUFM, ILUTE, and TLUMIP are disaggregate, i.e. apply microsimulation techniques. CUFM uses detailed land information in map form generated by a geographical information system. IRPUD starts with aggregate data but uses microsimulation in its housing market submodel; work is underway to make more submodels microscopic (Salomon et al., 2002). ILUTE and UrbanSim apply zones but use smaller spatial units such as grid cells or parcels in some submodels.

### *Dynamics*

The discussion on dynamics is related to the issue of equilibrium. Equilibrium models are based on the assumption that interdependent model variables, such as prices, supply, and demand, adjust to equilibrium with zero delay or, if adjustment is delayed, equilibrium is eventually reached. Dynamic models, on the other hand, are based on the assumption that some changes, e.g. changes in demand, are faster than others, e.g. responses of supply, and that these differences in speed of adjustment are so large that urban systems are normally in disequilibrium. All but three (BOYCE, IMREL, and KIM) of the 20 models are recursive simulation models. Recursive simulation models are called quasi-dynamic because, although they model the development of a city over time, within one simulation period they are in fact cross-sectional. This is, however, only true for strictly unified models. Composite models consist of several interlinked submodels that are processed sequentially or iteratively once or several times during a simulation period. This makes composite models well suited for taking account of time lags or delays due

to the complex superposition of slow and fast processes of urban development (Wegener et al., 1986). However, this feature is insufficiently used by some models, because their simulation period of 5 years has the effect of an implicit time lag – too long a time lag in most cases. This problem is likely to disappear as faster computers will make shorter simulation periods of 1 or 2 years more feasible.

### *Data requirements*

Data collection for a model of a large metropolis still requires major effort. However, in many cases the introduction of computers in local government has generated a pool of routinely collected and updated data that can be used as the information base for a model, in particular in the fields of population, housing, land use, and transport. Another factor reducing the data dependency of urban models is the significant progress made in urban theory in recent decades. The models of today are more parsimonious, i.e. can make do with fewer data than previous models. Examples illustrating this are the techniques to generate regional input–output matrices from national input–output matrices and regional totals through biproportional scaling methods; or techniques to create artificial microdata as samples from multivariate aggregate data.

### *Calibration and validation*

All 20 models of the sample have been (or could have been) calibrated using observed data, using readily available computer programs, and following well-established methods and standards. In particular, maximum-likelihood estimation of the ubiquitous logit model has become routine. Yet, while calibration has become easier, the limits to calibrating a model with data of the past have become visible. Calibration of cross-sectional models, as it is practiced today, provides the illusion of precision but does little to establish the credibility of models designed to look into the far future. There has been almost no progress in the methodology to calibrate dynamic or quasi-dynamic models. In the face of this dilemma, the insistence of some modelers on “estimating” every model equation appears almost an obsession. It would probably be more effective to concentrate instead on model validation, i.e. the comparison of model results with observed data over a longer period. In the future, the only real test of the performance of a model should be its ability to forecast the essential dynamics of the modeled system over a past period at least as long as the forecasting period.

### *Operationality*

All 20 models in the sample are operational in the sense that they have been applied to real cities. However, there are differences. Some models have

remained primarily research models applied to one particular study area. Others have been applied to only a few cities. Some are actually families of models, each specifically tailored to the needs of a particular urban area or client.

A few of the models are on their way to becoming standard software for a wider market. ITLUP has been used by a large number of metropolitan planning agencies in the USA. TRANUS stands out as a particularly advanced and well-documented software package, with an attractive user interface in Spanish or English. MEPLAN is being used in more and more cities all over the world, and DELTA in an increasing number of UK cities. The time seems not far when any planning office will be able to buy a complex and versatile urban model with full documentation, default values, and test data sets for less than a thousand dollars.

### *Applicability*

If one considers the enormous range of planning problems facing a typical metropolitan area in an industrialized country today, the spectrum of problems actually addressed with the 20 urban models in the sample is very narrow. The majority of applications answer traditional questions such as how land use planning or housing programs would affect land use development and transport, or how transport improvements or changes in travel costs would shift the distribution of activities in an urban area.

### **3. Future urban land use transport models**

Today there are many urban modeling projects underway all over the world. In the USA, environmental legislation, such as the Clean Air Act amendments of 1990, the Intermodal Surface Transportation Efficiency Act (ISTEA) of 1991, and the Transportation Equity Act for the 21st Century (TEA-21) of 1998, gave a boost to the development and application of urban land use transport models. ISTE A required cities to consider the likely effect of transportation policy decisions on land use development. In Europe, the European Commission has funded a number of studies employing land use transport models. The SPARTACUS project applied MEPLAN to three urban areas and connected the model with spatially disaggregate environmental submodels, the PROSPECTS project applied several models including DELTA and IMREL, and the PROPOLIS project applied MEPLAN, TRANUS, and IRPUD in seven urban regions in six European countries. There is an increasing number of applications of DELTA, MEPLAN, and TRANUS.

Nevertheless, there remain challenges to be met. The transport submodels used in most existing land use transport models do not apply state-of-the-art activity-based modeling techniques but the traditional four-step travel demand model

sequence, which is not suitable to model behavioral responses to many travel demand management policies presently discussed. Moreover, the spatial resolution of existing land use transport models is too coarse to model activity-based travel behavior or neighborhood-scale travel demand management policies.

Their insufficient spatial resolution is also one of the reasons why only very few land use transport models are linked to advanced environmental submodels of air quality, traffic noise, land take, and biotopes (Wegener, 1998a). Environmental issues are certain to play a more prominent role in the future, when the manifest unsustainability of present urban lifestyles and mobility patterns will increasingly come under scrutiny. However, most of the current efforts to link environmental submodels to transport or land use transport models are content to model emissions where actually air quality, i.e. local impacts of emissions occurring elsewhere, should be forecast.

This leads to issues of spatial equity. Most land use transport models are utilitarian in that they favor solutions yielding the greatest aggregate social benefit. However, urban societies are increasingly becoming socially and spatially fragmented and polarized, which means that distributional issues, both in social and spatial terms, are becoming more prominent. Distributional issues are particularly relevant in environmental conflicts, where polluters and those affected by pollution tend to come from different social groups or neighborhoods of a city. Most of the current land use transport models are insensitive to issues of social exclusion and spatial equity – one notable exception is the PROPOLIS project, in which different concepts of equity are explored with the results of land use transport models (LT et al., 2002).

The future of land use transport modeling will largely depend on whether emerging new models will live up to these challenges.

From a technical point of view, the prospects are excellent. More powerful computers will remove former barriers to increasing the spatial, temporal, and substantive resolution of models. The wealth of publicly available high-resolution spatial data will reduce aggregation error in spatial models. Geographic information systems (GIS) will become the mainstream data organization of urban models. Spatial disaggregation of land use and transport network data in raster GIS will permit the linkage between land use transport models and air quality and noise propagation models. Multiple representation of spatial data in raster and vector GIS will combine the advantages of spatial disaggregation (raster) and efficient network algorithms (vector). Aggregate probabilistic approaches (e.g. entropy maximizing) will be replaced by disaggregate stochastic (microsimulation) approaches.

Microsimulation was first used in social science applications by Orcutt et al. (1961), yet applications in a spatial context remained occasional experiments without deeper impact, though covering a wide range of phenomena such as spatial diffusion (Hägerstrand, 1968), urban development (Chapin and Weiss,

1968), transport behavior (Kreibich, 1979), demographic and household dynamics (Clarke, 1981), and housing choice (Kain and Apgar, 1985). Recently, though, microsimulation has seen renewed interest because of its flexibility to model processes that cannot be modeled in the aggregate (Clarke, 1996). In the last two decades, several microsimulation models of urban land use and transport have been developed (Landis, 1994; Wegener and Spiekermann, 1996; Landis and Zhang, 1998a,b; Waddell, 2002; Salomon et al., 2002).

A different approach emerged from the theory of cellular dynamics. Cellular automata (CA) are objects associated with areal units or cells. CA follow simple stimulus-response rules to change or not to change their state based on the state of adjacent or nearby cells. By adding random noise to the rules, surprisingly complex patterns that closely resemble real cities can be generated (Batty and Xie, 1994). More complex stimulus-response behavior is given to CA models in multi-reactive agents models. Multi-reactive agents are complex automata with the ability to control their interaction pattern; they can change not only their environment but also their own behavior, i.e. they are able to "learn" (Ferrand, 2000). The distinction between the behavior of multi-reactive agents and the choice behavior generated in microsimulation models is becoming smaller.

Probably the most advanced area of application of microsimulation in urban models is travel modeling (see Volume 1 in this series, *Handbook of Transport Modelling*). Aggregate travel models are unable to reproduce the complex spatial behavior of individuals and to respond to sophisticated travel demand management measures. As a reaction, disaggregate travel models aim at a one-to-one reproduction of spatial behavior by which individuals choose between mobility options in their pursuit of activities during a day (Axhausen and Gärling, 1992; Ben-Akiva et al., 1996). Activity-based travel models start from interdependent "activity programs" of household members of a "synthetic population" (Beckman et al., 1995) and translate these into home-based "tours" consisting of one or more trips. This way, interdependencies between the mobility behavior of household members and between the trips of a tour can be modeled as well as intermodal trips that cannot be handled in aggregate multimodal travel models. Activity-based travel models do not model peak-hour or all-day travel but disaggregate travel behavior by time of day, which permits the modeling of choice of departure time. There are also disaggregate traffic assignment models based on queuing or CA approaches, e.g. in the TRANSIMS project (Nagel et al., 1999), which reproduces the movement of vehicles in the road network with a level of detail not known before.

However, it will take some time until the first urban land use transport models fully based on microsimulation will be operational. Miller et al. (1998) presented a matrix in which the past and future evolution of urban land use transport models was charted. Figure 2 is an adaptation in which a sixth row (L6) has been added. In this figure the rows correspond to different levels of levels of land use modeling capability:

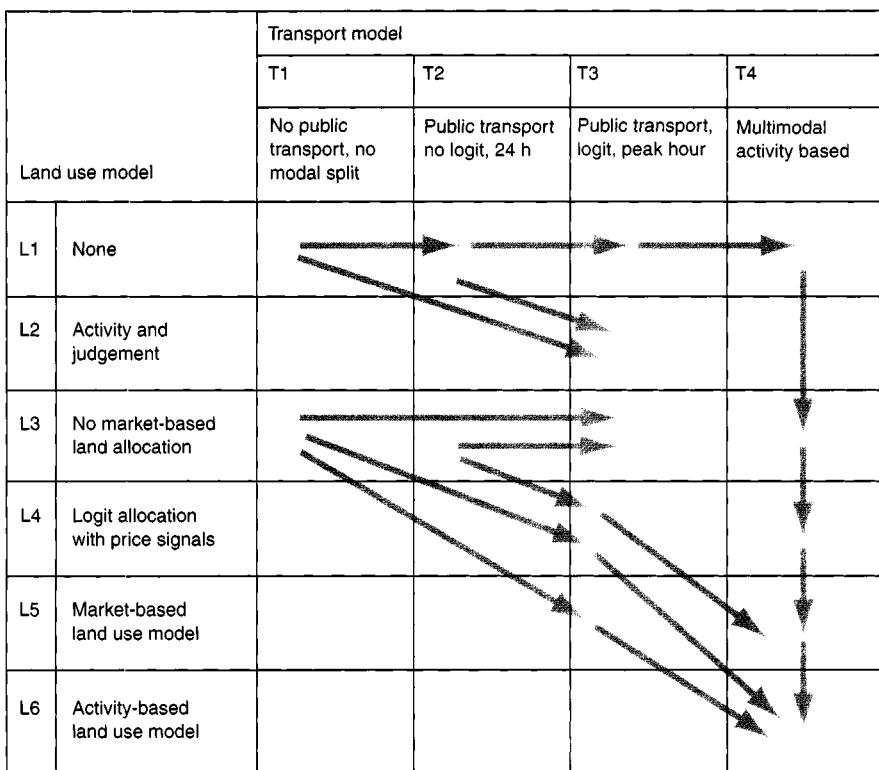


Figure 2. Evolution of urban land use transport models. (Adapted from Miller et al., 1998.)

- L1 – no land use model;
- L2 – activities are allocated to zones by professional judgement;
- L3 – non-market-based land allocation model;
- L4 – land allocation with price signals;
- L5 – fully integrated market-based model;
- L6 – activity-based land use model using microsimulation.

Similarly, the columns in Figure 2 represent different levels of travel demand modeling capability:

- T1 – only roads and automobile travel are modeled;
- T2 – public transport with simplified (non-logit) modal choice;
- T3 – logit-based modal choice, peak-period assignment;
- T4 – activity-based travel model using microsimulation.

Each cell in the figure therefore represents a land use transport modeling combination. The arrows indicate incremental paths that local governments can take to develop their land use transport modeling capability.

#### **4. Conclusions**

Predicting the impacts of integrated land use transport policies is a difficult task due to the multitude of concurrent changes of pertinent system variables. In general, there are three groups of methods to predict those impacts. The first one is to ask people about their anticipated reaction to changes such as increased transport costs or land use restrictions ("stated preference"). The second possibility is to draw conclusions from empirically observed behavior of people ("revealed preference"). The third group of methods comprises mathematical models to simulate human decision-making and its consequences. While all of the three possibilities have shortcomings, mathematical models are the only method able to forecast still unknown situations and to determine the effect of a single factor while keeping all other factors fixed.

Urban land use transport models incorporate the most essential processes of spatial development, including all types of land uses. Transport may be modeled either endogenously or by an exogenous transport model. Urban systems represented in land use transport models can be divided into nine subsystems according to the speed by which they change. The urban fabric consisting of infrastructure networks and land use patterns are subject to very slow change over time. Workplaces and housing change relatively slowly while the employment and residential population adjust their spatial behavior fairly quickly to changing circumstances. Goods transport or travel destinations are the most flexible phenomena of urban spatial development; they can be modified almost instantly according to changes in congestion or fluctuations in demand. There is a ninth subsystem, the urban environment, which is more complex regarding its temporal behavior.

A number of integrated land use transport models are in use today. There are significant variations among the models with respect to comprehensiveness, model structure, theoretical foundations, modeling techniques, dynamics, data requirements and calibration, and validation. Despite the achievements in developing these models further, there remain some challenges to be met. The transport submodels used in most current land use transport models do not apply state-of-the-art activity-based modeling techniques but rather the traditional four-step travel demand model sequence, which is inadequate for modeling behavioral responses to many currently applied travel demand management policies. The most promising technique for activity-based land use and transport

modeling is microsimulation, which makes it possible to reproduce the complex spatial behavior of individuals on a one-to-one basis.

In addition, the spatial resolution of present models is still too coarse to model neighborhood-scale policies and effects. In the future, the integration of environmental submodels for air quality, traffic noise, land take, and biotopes are likely to play a prominent role. Issues of spatial equity and socio-economic distributions are expected to gain similar importance in model building.

## Acknowledgment

The permission of Eric Miller and colleagues to use their matrix of model evolution as the basis for Figure 2 is gratefully acknowledged.

## References

- Alonso, W. (1964) *Location and land use*. Cambridge: Harvard University Press.
- Anas, A. (1983) "Discrete choice theory, information theory and the multinomial logit and gravity models," *Transportation Research B*, 17:13–23.
- Anas, A. (1994) *MetroSim: a unified economic model of transportation and land use*. Williamsville: Alex Anas.
- Anas, A. (1998) *NYMTC transportation models and data initiative: the NYMTC land use model*. Williamsville: Alex Anas.
- Anderstig, C. and L.-G. Mattsson (1998) "Modelling land use and transport interaction: evaluations and policy analysis," in: L. Lundqvist, L.-G. Mattsson and T.J. Kim, eds, *Network infrastructure and the urban environment: recent advances in land use/transportation modelling*. Berlin: Springer-Verlag.
- Axhausen, K.W. and T. Gärling (1992) "Activity-based approaches to travel analysis: conceptual frameworks, models and research problems," *Transport Reviews*, 12:324–341.
- Batty, M. (1994) "A chronicle of scientific planning: the Anglo-American modeling experience," *Journal of the American Planning Association*, 60:7–16.
- Batty, M. and Y. Xie (1994) "From cells to cities," *Environment and Planning B: Planning and Design*, 21:31–48.
- Beckman, R.J., K.A. Baggerly and M.D. McKay (1995) *Creating synthetic baseline populations*, LA-UR-95-1985. Los Alamos: Los Alamos National Laboratory.
- Ben-Akiva, M.E., J.L. Bowman and D. Gopinath (1996) "Travel demand model system for the information era," *Transportation*, 23:241–166.
- Berechman, J. and K.A. Small (1988) "Research policy and review 25: modeling land use and transportation: an interpretive review for growth areas," *Environment and Planning A*, 20:1283–1422.
- Boyce, D.E. (1988) "Renaissance of large-scale models," *Papers of the Regional Science Association*, 65:1–10.
- Boyce, D. and L.-G. Mattsson (1991) "Modeling residential location choice in relation to housing location and road tolls on congested urban highway networks," *Transportation Research B*, 33:581–591.
- Boyce, D.E., L.J. LeBlanc, K.S. Chon, Y.L. Lee and K.T. Lin (1981) *Combined models of location, destination, mode and route choice: a unified approach using nested entropy constraints*. Publication No. 3. Urbana: Transportation Planning Group, Department of Civil Engineering, University of Illinois.

- Boyce, D.E., K.S. Chon, Y.J. Lee, K.T. Lin and L. LeBlanc (1983) "Implementation and computational issues for combined models of location, destination, mode, and route choice," *Environment and Planning A*, 15:1219–1230.
- Brotchie, J.F., J.W. Dickey and T. Sharpe (1980) *TOPAZ planning techniques and applications. Lecture notes in economics and mathematical systems series*, Vol. 180. Berlin: Springer-Verlag.
- Chapin, F.S. and S.F. Weiss (1968) "A probabilistic model for residential growth," *Transportation Research*, 2:375–390.
- Clarke, G.P., ed. (1996) *Microsimulation for urban and regional policy analysis. European Research in Regional Science* 6. London: Pion.
- Clarke, M. (1981) "A first-principle approach to modelling socio-economic interdependence using microsimulation," *Computers, Environment and Urban Systems*, 6:211–227.
- de la Barra, T. (1989) *Integrated land use and transport modelling*. Cambridge: Cambridge University Press.
- de la Barra, T. (1998) "Improved logit formulations for integrated land use, transport and environmental models," in: L. Lundqvist, L.-G. Mattsson and T.J. Kim, eds, *Network infrastructure and the urban environment: recent advances in land use/transportation modelling*. Berlin: Springer-Verlag.
- Domencich, T.A. and D. McFadden (1975) *Urban travel demand: a behavioral analysis*. Amsterdam: North Holland.
- Echenique, M.H., A.D.J. Flowerdew, J.D. Hunt, T.R. Mayo, I.J. Skidmore and D.C. Simmonds (1990) "The MEPLAN models of Bilbao, Leeds and Dortmund," *Transport Reviews*, 10:309–322.
- Ferrand, N. (2000) "Multi-reactive agents paradigm for spatial modelling," in: A.S. Fotheringham and M. Wegener, eds, *Spatial models and GIS: new potential and new models*. London: Taylor and Francis.
- Haag, G. (1990) "Master equations," in: C.S. Bertuglia, G. Leonardi and A.G. Wilson, eds, *Urban dynamics. Designing an integrated model*. London: Routledge.
- Hägerstrand, T. (1968) *Innovation diffusion as spatial process*. Chicago: University of Chicago Press.
- Hansen, W.G. (1959) "How accessibility shapes land use," *Journal of the American Institute of Planners*, 25:73–76.
- Harris, B. (1985) "Urban simulation models in regional science," *Journal of Regional Science*, 25:545–567.
- Harris, B. (1994) "Science in planning: past, present, future," *Journal of the American Planning Association*, 60:31–34.
- Hensher, D. and T. Ton (2001) "TRESIS: a transportation, land use and environmental strategy impact simulator for urban areas," in: *8th World Conference on Transport Research*, Paper. Seoul.
- Hunt, J.D. and J.E. Abraham (2003) "Design and application of the PECAS land use modelling system," in: *8th International Conference on Computers in Urban Planning and Urban Management*, Paper. Sendai.
- Hunt, J.D. and D.C. Simmonds, (1993) "Theory and application of an integrated land use and transport modelling framework," *Environment and Planning B: Planning and Design*, 20:221–244.
- Hutchinson, B., P. Nijkamp and M. Batty, eds (1985) *Optimization and discrete choice in urban systems*. Berlin: Springer-Verlag.
- Kain, J.F. and W.C. Apgar, Jr (1985) *Housing and neighborhood dynamics: a simulation study*. Cambridge: Harvard University Press.
- Kim, T.J. (1989) *Integrated urban systems modeling: theory and applications*. Dordrecht: Kluwer.
- Kreibich, V. (1979) "Modelling car availability, modal split, and trip distribution by Monte Carlo simulation: a short way to integrated models," *Transportation*, 8:153–166.
- Landis, J.D. (1994) "The California urban futures model: a new generation of metropolitan simulation models," *Environment and Planning B: Planning and Design*, 21:399–422.
- Landis, J.D. and M. Zhang (1998a) "The second generation of the California urban futures model. Part 1: model logic and theory," *Environment and Planning B: Planning and Design*, 25:657–666.
- Landis, J.D. and M. Zhang (1998b) "The second generation of the California urban futures model. Part 2: specification and calibration results of the land use change submodel," *Environment and Planning B: Planning and Design*, 25:795–824.
- Lowry, I.S. (1964) *A model of metropolis*, RM-4035-RC. Santa Monica: Rand.
- LT, IRPUD, ME&P, MECSA, STRATEC, TRT and UCL (2002) *PROPOLIS (Planning and Research of Policies for Land Use and Transport for Increasing Urban Sustainability)*. Helsinki: LT Consultants (<http://www.ltcon.fi/propolis>).

- Mackett, R.L. (1985) "Integrated land use transport models," *Transport Reviews*, 5:325–343.
- Mackett, R.L. (1991a) "A model-based analysis of transport and land use policies for Tokyo," *Transport Reviews*, 11:1–18.
- Mackett, R.L. (1991b) "LILT and MEPLAN: a comparative analysis of land use and transport policies for Leeds," *Transport Reviews*, 11:131–54.
- Martinez, F.J. (1991) "Transport investments and land values interaction: the case of Santiago City," in: *Proceedings of the PTRC Summer Annual Meeting*. London: PTRC.
- Martinez, F.J. (1992) "The bid-choice land use model: an integrated economic framework," *Environment and Planning A*, 24:871–885.
- Martinez, F.J. (1996) "Analysis of urban environmental policies assisted by behavioural modelling," in: Y. Hayashi and J. Roy, eds, *Transport, land use and the environment*. Dordrecht: Kluwer.
- Martinez, F.J. and P.P. Donoso (1995) "MUSSA model: the theoretical framework," in: D.A. Hensher and J. King, eds, *World Transport Research. Proceedings of the 7th World Conference on Transportation Research*, Vol. 2. Oxford: Pergamon.
- Miller, E.J., D.S. Kriger, J.D. Hunt and D.A. Badoe (1998) *Integrated urban models for simulation of transit and land use policies*, Final Report, TCRP Project H-12. Toronto: Joint Program of Transportation, University of Toronto.
- Miyamoto, K. and R. Udomsri (1996) "An analysis system for integrated policy measures regarding land use, transport and the environment in a metropolis," in: Y. Hayashi and J. Roy, eds, *Transport, land use and the environment*. Dordrecht: Kluwer.
- Nagel, K., R.J. Beckman and C.L. Barrett (1999) *TRANSIMS for transportation planning*, LA-UR 98-4389. Los Alamos: Los Alamos National Laboratory ([http://transims.tsasa.lanl.gov/PDF\\_Files/LAUR98-4389.pdf](http://transims.tsasa.lanl.gov/PDF_Files/LAUR98-4389.pdf)).
- Orcutt, G., M. Greenberger, A. Rivlin and J. Korbel (1961) *Microanalysis of socioeconomic systems: a simulation study*. New York: Harper and Row.
- Oregon Department of Transportation (2002) <http://www.odot.state.or.us/tddtpau/modeling.html>.
- Pfaffenbichler, P.C. and S.P. Shepherd (2002) "A dynamic model to appraise strategic land use and transport policies," *European Journal of Transport and Infrastructure Research*, 2:255–283.
- Prastacos, P. (1986) "An integrated land use-transportation model for the San Francisco region," *Environment and Planning A*, 18:307–322, 511–528.
- Putman, S.H. (1991) *Integrated urban models 2. New Research and applications of optimization and dynamics*. London: Pion.
- Putman, S.H. (1998) "Results from implementation of integrated transportation and land use models in metropolitan regions," in: L. Lundqvist, L.-G. Mattsson and T.J. Kim, eds, *Network infrastructure and the urban environment: recent advances in land use/transportation modelling*. Berlin: Springer-Verlag.
- Rho, J.H. and T.J. Kim (1989) "Solving a three-dimensional urban activity model of land use intensity and transport congestion," *Journal of Regional Science*, 29:595–613.
- Salomon, I., P. Waddell and M. Wegener (2002) "Sustainable life styles? Microsimulation of household formation, housing choice and travel behaviour," in: W.R. Black and P. Nijkamp, eds, *Social change and sustainable transport*. Bloomington: Indiana University Press.
- Simmonds D.C. (2001) "The objectives and design of a new land use modelling package: DELTA," in: G.P. Clarke and M. Madden, eds, *Regional science in business*. Berlin: Springer-Verlag.
- Waddell, P. (2002) "UrbanSim: modeling urban development for land use, transportation and environmental planning," *Journal of the American Planning Association*, 68:297–314.
- Wegener, M. (1982a) "A multilevel economic-demographic model for the Dortmund region," *Sistemi Urbani*, 3:371–401.
- Wegener, M. (1982b) "Modeling urban decline: a multilevel economic-demographic model of the Dortmund region," *International Regional Science Review*, 7:21–41.
- Wegener, M. (1986) "Transport network equilibrium and regional deconcentration," *Environment and Planning A*, 18:437–56.
- Wegener, M. (1994) "Operational urban models: state of the art," *Journal of the American Planning Association*, 60:17–29.
- Wegener, M. (1995) "Current and future land use models," in: G.A. Shunk, P.L. Bass, C.A. Weatherby and L.J. Engelke, eds, *Travel Model Improvement Program Land Use Modeling Conference Proceedings*. Washington, DC: US Department of Transportation.

- Wegener, M. (1996) "Reduction of CO<sub>2</sub> emissions of transport by reorganisation of urban activities," in: Y. Hayashi and J. Roy, eds, *Transport, land use and the environment*. Dordrecht: Kluwer.
- Wegener, M. (1998a) "Applied models of urban land use, transport and environment: state-of-the-art and future developments," in: L. Lundqvist, L.-G. Mattsson and T.J. Kim, eds, *Network infrastructure and the urban environment: recent advances in land use/transportation modelling*. Berlin: Springer-Verlag.
- Wegener, M. (1998b) *The IRPUD model: overview*. Dortmund: Institute of Spatial Planning, University of Dortmund (<http://irpud.raumplanung.uni-dortmund.de/irpud/pro/mod/mod.htm>).
- Wegener, M. and F. Fürst (1999) *Land use transport interaction: state of the art*. Berichte aus dem Institut für Raumplanung 46. Dortmund: Institute of Spatial Planning, University of Dortmund (<http://www.inro.tno.nl/transland/Deliverable%202a.pdf>).
- Wegener, M. and K. Speckermann, (1996) "The potential of microsimulation for urban models," in: G. Clarke, ed., *Microsimulation for urban and regional policy analysis. European Research in Regional Science 6*. London: Pion.
- Wegener, M., F. Gnäd and M. Vannahme (1986) "The time scale of urban change," in: B. Hutchinson and M. Batty, eds, *Advances in urban systems modelling*. Amsterdam: North Holland.
- Wegener, M., R.L. Mackett and D.C. Simmonds (1991) "One city, three models: comparison of land use/transport policy simulation models for Dortmund," *Transport Reviews*, 11:107–129.
- Wilson, A.G. (1967) "A statistical theory of spatial distribution models," *Transportation Research*, 1:253–269.
- Wilson, A.G. (1997) "Land use/transport interaction models – past and future," *Journal of Transport Economics and Policy*, 32:3–23.

*Chapter 10*

## INTEGRATED LAND USE/TRANSPORT MODEL REQUIREMENTS

ERIC J. MILLER

*University of Toronto*

### 1. Introduction

The case for using integrated land use/transport models in urban transport analysis is strong (e.g. see Hensher, 2002a). There is growing interest in developing and applying integrated models in planning contexts, and a number of operational integrated models exist worldwide. Some of the more commonly known of these models include MEPLAN (Hunt and Simmonds, 1993), TRANUS (de la Barra, 1989), MetroSim (Anas, 1995), MUSSA (Martinez and Donoso, 2001), UrbanSim (Waddell, 2000), TRESIS (Hensher, 2002b) and IRPUD (Wegener, 2000), among others. General reviews of the integrated modeling state of practice include Southworth (1995), Wegener (1995), and Miller et al. (1998).

This chapter discusses the general design requirements of integrated models of urban land use/transportation, with at least two purposes in mind. The first is to provide a general specification for integrated models, within which specific modeling approaches discussed in other chapters of this handbook can be discussed and compared. Second, the design issues and criteria discussed in this chapter provide a basis for evaluating alternative operational models that might be considered for implementation within a given planning context.

To address these objectives, the chapter begins with the definition of a generic, idealized modeling system that is sufficiently general to serve as the framework for considering virtually any integrated model. The chapter then discusses a set of design issues that need to be addressed in developing an integrated model, followed by definition of criteria that can be used in evaluating alternative models.

### 2. A framework for integrated modeling

Figure 1 presents a highly idealized representation of a comprehensive land use/transportation modeling system. The “behavioral core” of this system (shaded area of Figure 1) consists of four inter-related components:

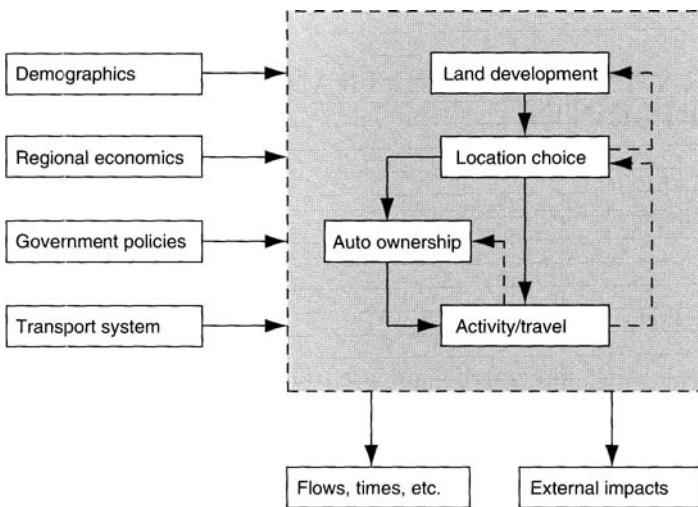


Figure 1. Integrated urban modeling system framework.

- Land development – this models the evolution of the built environment, and includes both the initial development of previously “vacant” land and the redevelopment over time of existing land uses. This component could also be labeled “building supply,” since building stock supply functions (construction, demolition, renovation, etc.) are included.
- Location choice – this includes the location choices of households (for residential dwellings), firms (for commercial locations), and workers (for jobs).
- Activity/travel – whether performed by traditional four-stage methods or by emerging activity-based models, this component involves predicting the trip-making behavior of the population, ultimately expressed in terms of origin-destination flows by mode by time of day. It also, ideally, includes the prediction of goods and services movements associated with the functioning of the urban economic system.
- Auto ownership – this component models household auto ownership levels – an important determinant of household travel behavior.

In speaking about “land use models” it is common for transportation planners to blur the distinctions among these four components, especially between the concepts of land development and location choice. A properly specified model, however, should clearly distinguish among these components since they involve very different actors, decision processes, and time-frames. They also represent distinctly different “degrees of freedom” for the system to respond to policy inputs.

Each component in Figure 1 involves a complex set of submodels. In particular, market-based supply–demand relationships tend to dominate aggregate behavior in each case (buyers and sellers of houses interact within the housing market; workers and employers interact within the labor market; etc.), with prices (or, in the case of trip making, travel time) both being determined by and playing a major role in determining the outcome of these supply–demand interactions. Models that ignore these major supply–demand interactions may fail to properly capture the dynamic evolution of the urban system over time.

A simple flowchart such as Figure 1 never properly captures the temporal complexities of a dynamic system. The vertical hierarchy is chosen to represent short-run conditioning effects. That is, in the short run, most location choices are made within a “fixed” building stock supply. Similarly, in the short run, most activity/travel decisions are made given a “fixed” distribution of activity locations (and a fixed number of household autos). In the longer run, all four components evolve, at least partially in response to “feedback” from lower levels in the hierarchy (land use evolves in response to location needs of households and firms; people relocate their homes and/or jobs at least partially in response to accessibility factors; etc.).

The inclusion of auto ownership as a separate box within the “behavioral core” is somewhat unconventional. Auto ownership is often treated as simply one more input to travel models. As Ben-Akiva (1974) has observed, however, auto ownership is an integral part of the “mobility bundle” (which, in terms of Figure 1, Ben-Akiva would define as the combination of the location choice, auto ownership, and activity/travel components) in that it is fundamentally interconnected with residential location and work trip commuting decision-making. This point is strongly reinforced within the empirical literature (e.g. see Badoe and Miller, 2000), in which auto ownership is consistently found to be an important “intermediate variable” connecting urban form with travel behavior.

As shown in Figure 1, there are at least four major “drivers” of urban systems:

- demographics – evolution of the resident population in terms of its age–sex distribution, population size, education level, household composition, etc.;
- regional economics – evolution of the urban region economy in terms of its size, industrial distribution, etc.;
- government policies – zoning, taxation, interest rates, etc.;
- the transportation system – road, transit, non-motorized, etc.

The extent to which these various drivers are explicitly included within the model will vary from one modeling system to another. Government policies and changes to the transportation system are almost exclusively treated as model inputs; demographic and regional economic processes are almost always at least partially included within the modeling system. The key point is that the full range of “drivers” of land use/location/travel decision-making should be included in the

**Box 1**  
**Integrated urban model design issues**

**Physical system representation**

- Time
- Space (land)
- Building stock
- Transportation networks
- Services

**Representation of processes**

- Land development
- Location choices
- Job market
- Demographics
- Regional economics
- Automobile holdings
- Activity/travel demand
- Network performance

**Representation of decision-makers**

- Persons
- Households
- Private firms
- Public authorities

**"Generic issues"**

- Level of aggregation/disaggregation
- Endogenous versus exogenous treatment
- Level of "process type"
- Model specification

**Implementation issues**

- Data requirements
- Computational requirements
- Technical support requirements

modeling system to ensure that the impact of any one policy can be properly represented and evaluated by the model. In particular, it was often the case with early land use models that they overemphasized transportation system effects on land use/location processes, and hence were biased toward over-predicting the impact of transportation system improvements on these processes. Transportation improvements, however, are only one among many determinants of land development decisions (e.g. see Knight and Trygg (1977) for a classic discussion of the role of rail transit on land development).

It is fair to say that no existing transportation/land use model fully captures all aspects of the comprehensive modeling system sketched above. It does, however, define the goal toward which all integrated urban models should be striving, in that such a system would provide the analytical means for assessing the short and long run impacts of transport alternatives in a balanced and comprehensive way.

### 3. Design issues

A large number of issues must be considered in the design of an operational integrated urban model derived from the idealized modeling system presented above. A number of these are listed in Box 1 and briefly discussed below. Different models, of course, will address these issues in a variety of ways, ranging from ignoring them completely to dealing with them in a very computationally detailed and/or theoretically rigorous manner. No “right” answer/approach necessarily exists with respect to any one of these issues. As with any design exercise, the “right” or “best” design depends on the specific application context (data availability, computational and technical support capabilities, analysis/forecasting needs, size of the urban area, etc.). In addition, no one issue or “dimension” of the problem can be “optimized” in isolation; it is the overall balance across design dimensions that is important (e.g. very fine spatial resolutions may be difficult/impossible/unnecessary to maintain within very long-range forecasting applications). The intention here, rather, is simply to generate a reasonably comprehensive list of issues that should be considered in developing or implementing an integrated model.

The identified design issues have been grouped in Box 1 into five categories. The first three categories deal with the substance of the system being modeled: the physical entities, the behavioral entities, and the processes by which these physical and behavioral entities evolve over time. The last two categories are more methodological in nature, dealing with how the representation of these entities and processes is actually implemented within an operational modeling system. Each of these groups of issues are discussed in turn below.

#### 3.1. Physical system representation

Fundamental to model design are decisions concerning the representation of the physical elements of the system: time, land (space), buildings, transportation networks, and other physical infrastructure. These decisions fundamentally affect the precision and accuracy of the model, its data and computational requirements, and options for the representation of behavior within the physical urban system.

##### *Treatment of time*

All forecasting models must predict how an urban system state in some base year is likely to evolve into the future, typically up to some user-specified forecast “horizon year.” Choices of model base and horizon years, and the time increment or step used to move the system from the base to horizon year, are fundamental design questions.

Also fundamental is the treatment of “dynamics” within the model. Many models assume that system equilibrium is achieved in each time step, and so are able to appeal to the mathematical conditions for equilibrium to solve for the system state at the end of each time step. Ordinary gravity models are a classic example of this approach. Other models do not assume equilibrium. Rather, they explicitly simulate the evolution of the system state from one point in time to another as a function of various assumed processes.

The question of system dynamics is further complicated by the fact that different processes at work within the urban system operate on different time-frames. Land development processes operate over time periods of decades or more; many household-level decisions are perhaps made on approximately a yearly basis (e.g. residential relocation decisions, automobile transaction decisions, and household structure evolution); many activity/travel decisions change from week to week and from day to day; road network operating conditions (and, hence, energy consumption and tailpipe emissions) vary from minute to minute and second to second. Reconciling this wide combination of “slow” (or long-run) and “fast” (or short-run) dynamics within an overall modeling system is challenging, to say the least.

### *Treatment of space*

The spatial nature of urban systems represents one of the major sources of complexity in the analysis and modeling of these systems. Space enters both in terms of the locations of activities and in terms of the flows of people, goods, etc., between these activity locations. Design issues include zone system definition, degree of use of/interface with geographic information system (GIS) software, and the degree to which “micro” neighborhood design attributes are incorporated into the set of spatial attributes maintained within the model. Modern GIS create the potential for non-zone-based models. Even with such capabilities, however, it is likely that a zone system is required, at a minimum for data display purposes. Certainly all current and immediately foreseeable integrated models depend on a spatial zone system.

### *Building stock*

While we often talk rather loosely about “land use,” most urban activities actually occur within buildings of one type or another, and the built environment, to a large extent, determines the nature of which activities occur where. The extent to which building stock (by amount, type, etc.) is explicitly represented within the model represents an important design decision, and is found to vary considerably from one model to another.

### *Transportation networks*

Appropriate representation of both road and transit systems is clearly an essential component of any integrated urban model. Issues here include maintaining consistency in level of detail with the zone system being used; appropriate representation of transit walk access/egress; and appropriate representation of parking supply.

Central to the design of the transportation network representation is the nature of the network performance and route assignment model to be used. In recent years a wide variety of dynamic, often microsimulation-based models of network performance and route choice have been developed, primarily for real-time network control or other intelligent transportation systems related applications. Such models, however, are computationally extremely intensive and require a very finely detailed network representation. It is simply not clear, however, the extent to which this level of detail (and associated computational burden) is required or even feasibly supportable within an integrated urban model that is typically intended for medium- to long-term forecasting applications.

### *Services*

The development of land depends on the provision of many services in addition to transportation. These include sewers; water; electricity/gas; communications (telecommunications, fiber optic networks, etc.); heating/cooling; and proximity to emergency services (fire stations, hospitals, etc.). Explicit treatment of non-transport services and infrastructure varies among modeling systems, but generally explicit representation of such services is minimal at best. Ideally, however, the model architecture should be extensible to include a broader and more detailed representation of services as time, opportunity and need warrant.

### *3.2. Representation of active agents*

“Active agents” are the decision-making units – the people, households, firms, etc., who actually cause the urban area to exist and to evolve over time, through their various activities. People buy and sell homes; participate in the labor market; travel every day to and from work, school, shopping, etc.; get married and (sometimes) have babies; age and (eventually) die; etc. Each individual lives within either a family or non-family unit generally referred to as a household. For many important activities, such as residential location choice and automobile holdings choice, the household is in most cases the natural decision-making unit, rather than the individual. Thus, the possibility exists that one might wish to

explicitly represent both individual persons and households as interrelated but identifiably separable decision-making units within the model.

Firms similarly face location-relocation decisions and go through a life-cycle process of "birth," growth, and, often, "death." Private firms provide the majority of the "economic energy" fueling urban processes. They occupy land and buildings; they demand and supply goods and services; they employ workers. Two types of firm that are of particular interest within the model are developers (whose business it is to develop land and construct buildings) and transportation firms (whose business it is to provide transportation services to themselves and/or others).

Other active agents obviously exist within urban areas that have direct impacts on the land use/transport interaction, notably various government agencies. The extent to which such agents are explicitly incorporated within an integrated urban model is another design decision, although, in general, they are usually assumed to act externally to the processes being explicitly modeled. In particular, much of the behavior of public authorities lies outside the domain of the model *per se* in that it represents the political and bureaucratic processes of public policy debate and decision-making, the outcomes of which become policy inputs to the model. At the same time, however, public authorities are typically major employers and consumers (and providers) of land, floor space, transportation, and other goods and services within urban areas, and so have roles to play within the model as well.

### *3.3. Representation of processes*

Box 1 lists the primary processes that collectively define the transportation-land use interaction. Most of these processes have already been discussed in the previous section. Additional points to note concerning process representation within integrated urban models include the following.

#### *Market processes*

As has already been observed, many of these processes are market-driven. These include land development/building supply, residential and commercial real estate markets, labor markets, and travel markets. "Proper" representation of both demand and supply processes within each of these markets is essential to modeling such processes successfully. Implicit in this observation is that prices should be explicitly represented within the model and should be determined within the model through the demand-supply interaction.

#### *Network performance*

Network model design issues include choice of static versus dynamic route assignment procedures, deterministic versus stochastic procedures, equilibrium

versus non-equilibrium assumptions, level of network detail, compatibility with/support for emissions and energy use models, and degree of integration between road and transit network representation and processes.

### *Regional economics*

Regional economics is shown as one of the major “drivers” of the urban system in Figure 1, with the implication that it lies outside of the modeling system *per se*. Given that the magnitude and nature of the economic activity that occurs within an urban area depends in no small way on regional economic factors, some integrated urban models incorporate input–output models of the local and regional economies as a major component of the overall modeling system. Indeed, in the limit, the entire modeling system can be developed as a model of a spatially distributed economic system, with the consumption of land, buildings and travel as being but three of the many economic sectors being modeled.

### *Demographics*

While shown as an external “driver” of the urban system in Figure 1, demographic processes are, in fact, an integral part of the urban system and its internal dynamics. Births, deaths, aging, and household formation/evolution/dissolution are fundamental processes determining the characteristics of the population and thereby the demand for housing, education, jobs, goods, services, etc. Much of the explanatory power of disaggregate models of human decision-making comes from being able to specify the attributes of the individuals involved, and hence to be able to say something with reasonable confidence about their tastes and preferences, etc. If such disaggregate decision models are to be employed effectively, then the overall modeling system must be able to supply these decision models with the required decision-maker attributes. As a result, a central and significant component of an integrated modeling system is, ideally, a strong, dynamic model of person and household demographics. This includes both the capability to synthesize, if need be, the attributes of individuals and households in the base (initial) system state, and to “update” or “evolve” these attributes over time within the overall simulation run.

Different models will, of course, deal with these processes in various ways, including implicitly through the combination of two or more processes within a single submodel. For example, a simple Lowry-type model, in which the residential population is allocated over the urban area using a logit or gravity model (given known employment locations) effectively combines residential land development, residential location choice, and job location choice into a single “net” or “reduced-form” model. One way or another, however, every integrated model must deal with each of the processes listed in Box 1.

### *3.4. Generic design issues*

Integral to the design of the representation of the physical system, the behavioral agents and the processes at work within the system are fundamental choices concerning aggregation level, boundaries between what is endogenous to the model and what is not, and “process type.” Each of these is briefly discussed below.

#### *Level of aggregation/disaggregation*

Most currently operational integrated models are quite aggregate in both space and time, often using less than 100 zones to represent an entire urban area and working in time steps of 5 or even 10 years. At the other extreme, many researchers are experimenting with “microsimulation” models, in which individual households, building, firms, etc., are the basic model building blocks. Choice of aggregation level will have profound effects on data requirements, options for modeling processes, computation requirements, etc., and represents one of the primary, distinguishing decisions in any model design.

We are most used to thinking of the aggregation issue in terms of spatial aggregation (i.e. use of zones instead of individual people as the unit of analysis; size of zones used; etc.). Aggregation decisions, however, are made with respect to every entity (physical or behavioral) and every process included in the model. Use of a 5 year time step to represent a process that occurs on a yearly (or shorter) basis constitutes temporal aggregation. Not including potentially salient personal attributes (say, for example, education level or occupation type) in decision-making models represents aggregation over “attribute space.” And so on.

#### *Endogenous versus exogenous factors*

Any agent or process whose attributes and/or behavior are determined within the model is said to be endogenous to the model. Conversely, factors that affect system performance but whose values are simply provided to the model as inputs are called exogenous factors. A fundamental step in any model design involves “drawing the boundaries” around the model; that is, determining what is to be included within the model versus what will be excluded. As with the aggregation discussed above, these decisions will directly affect data and computing requirements, policy sensitivity, and process-modeling options.

#### *Process type*

Decisions must be made concerning how to model each endogenous process within the model. While a near-continuum of options exist, these can be broadly defined as falling into two categories: “transition models” and “choice models”

(Wegener, 1995). Transition models use simple deterministic or probabilistic rules for determining changes in attributes, system states, etc., over time. Examples of transition models include most models for most demographic processes, such as deterministic population aging models (e.g. add 1 year to each person's age for each year being simulated) and fertility models that express the probability of a woman giving birth to a child as a simple function of her age, marital status, etc. Choice models, on the other hand, attempt to model explicitly the choice process underlying a particular decision or action (random utility choice models and computational process models are both obvious examples of this class of model). Residential location choice, employment location choice, auto ownership, and activity/travel decisions are all examples of processes that might typically be modeled as choice processes within an integrated urban model.

While some processes may "obviously" fall into one category or the other (aging is a pure transition process – we have no choice in the matter whatsoever!), allocation of a given process to one type of modeling approach or the other is at least partially dependent on the application context, available data and modeling methods, computational resources, etc. Household formation and evolution, for example, in "real life" certainly are the result of complex interpersonal decision-making. In most integrated urban models, however, such processes (if endogenously modeled at all) are represented using relatively simple transition models.

### *Model specification*

This includes both the selection of model functional form (logit model, etc.) and the explanatory variables to be included within the model. This issue is so integral to all model building that there is perhaps little that needs to be said with respect to it, except to point out the obvious facts that model specification determines theoretical soundness (and hence the fundamental credibility of the model), computational intensity, data requirements, and policy sensitivity (if a particular policy-relevant variable is not included in the model, then the model obviously will not be able to respond to the given policy).

### *3.5. Implementation issues*

All models require data, computational resources, and technical support to be developed, implemented, and maintained as an operational tool. Each of these issues is briefly discussed below.

#### *Data requirements*

Historical data are required for both model estimation/calibration and validation. Estimation usually refers to the statistical determination of model parameters

that cause the model to "best fit" (in a statistically well-defined sense) to observed, historical data (e.g. use of maximum likelihood estimation to estimate logit choice model parameters, or use of linear regression analysis to estimate trip generation model parameters). Calibration usually refers to post-estimation "parameter adjustments" that "force" the model to better replicate observed data (e.g. use of  $K$  factors in gravity trip distribution models to force the model to reproduce observed screenline or cordon counts). Given the complexity of most integrated urban models (typically involving many submodels, each one of which possessing its own level of complexity, often exercised within a simulation framework), a considerable amount of calibration as opposed to estimation is usually required in order to get these models "working properly." This, in turn, implies the need for considerable experience and good professional judgement to be applied to the model development process.

Once a model has been estimated/calibrated, it should be validated as a forecasting tool by performing "historical forecasts" between two or more points in time in the past for which historical data are available. For example, a model might be calibrated using data from 1990 and 1995. Using 1995 as a base, it then may be used to "forecast" year 2000 conditions. This 2000 "forecast" can then be compared with known data for 2000 in order to assess the ability of the model to predict beyond the time period covered by the calibration data.

The foregoing discussion indicates that integrated urban models typically require a considerable amount of historical data from multiple time periods in order to be calibrated and validated. Typically, at least three time periods are required: two for model estimation/calibration, and a third for validation. The likely availability of historical data (what variables at what level of spatial detail for what years at what level of reliability, etc.) must be considered in the model design process, since there is no use in designing a modeling system that cannot be implemented due to data restrictions. Known, insurmountable data limitations will often drive the model design with respect to such important factors as time step, level of spatial aggregation, and choice of model specification.

Once a model is operational, it requires a new type of data to be used as a forecasting tool: estimated values of the exogenous inputs to the model for future years being simulated by the model. These estimates may come from policy scenarios, professional judgement, other models, etc., but, one way or another, they must be provided by the analyst to the model so that it can be run. These input data can be quite extensive, difficult to generate, and, of course, subject to error. In general, a classic trade-off exists in model design between "specification error" (which is built into the model due to model simplifications, abstractions, etc., that cause the model to fail to perfectly capture real world behavior) and "forecast error" (error introduced during the forecasting process by inaccurate inputs). As with the model development data requirements, the forecast input data requirements must also be considered during the model design process, and,

again, may well impose significant practical constraints on model design with respect to the temporal, spatial and/or behavioral representations that are feasible to achieve.

Integrated urban models are well known to be extremely “data hungry.” At any point in time, data availability may well prove to be the single biggest constraint on model design and application. At least two more positive observations, however, with respect to data are the following.

- The datasets available to support integrated urban modeling have improved dramatically over the last twenty-five years, and can be expected to continue to improve as we move even more deeply into the “information age.” The contributions of GIS to our ability to store and make effective use of large spatial databases are particularly noteworthy in this regard.
- The need for improved datasets to support integrated modeling can prove to be a very positive stimulus for improving the overall planning database in urban areas. That is, while perhaps initially motivated by modeling requirements, once collected and assembled, databases, if properly managed, can take on a life of their own and can provide very useful support for a range of planning applications. That is, a good historical database can support a wide range of historical analyses that result in improved understandings of processes and issues at work within the urban area; that is the database can and should support various descriptive and diagnostic analyses as well as model-building and forecasting applications.

### *Computational requirements*

Integrated urban models are by necessity computer based. The size of the computer (CPU, memory, disk space, etc.) required to house the model, the time required to execute a single run of the model (with obvious trade-offs between run time and computer size), and the software required to implement and support the model (i.e. the actual computer code within which the model is implemented, as well as the ancillary software – operating system, GIS, database management system, statistical analysis systems, etc.) are all of critical concern within the model design process. Historically, the computing power cost-effectively available to researchers and planners has imposed significant limitations on the scale and scope of integrated urban models. Past and continuing advances in computer technology, however, are fast removing these barriers. The amazing power of desktop computers, the continuing emergence of parallel processing, the explosion of commercially available software, etc., are all extending the boundaries of what is feasible, to the point that computing power *per se* is probably no longer the constraint on practical modeling systems that it once was.

### *Technical support requirements*

The discussion to this point has focused on the model design and development process. Implementation of a model within a given planning agency, and then the ongoing maintenance and use of the model within this agency, requires significant technical support. In-house staff must be dedicated to the operation of the model, have appropriate professional backgrounds and have been properly trained in the understanding and use of the model. An institutional, management-level commitment must exist within the planning agency to provide the time, money, and moral support required to get the model implemented and then to keep it operating effectively and efficiently. And adequate and ongoing support must also be available from the model developers (who usually will be external to the planning agency) with respect to training, trouble shooting, and ongoing system maintenance and upgrading. While largely implementation and operations, rather than design, oriented, the design implication of this issue is that an overly complex model design that is difficult to understand, operate, and maintain, or that is not "robust" with respect to its ease of use within an operational planning environment, will not be an attractive or even practical model for application within such contexts.

## **4. Evaluation criteria**

Box 2 lists a set of evaluation criteria that can be used in the assessment of any operational integrated model. While these criteria obviously relate directly to the design issues discussed in the previous section, they also represent a different (and generally somewhat more abstract) "slice" through the problem. These criteria have been divided into three groups, relating to the credibility of the models for use within operational planning applications; the feasibility of implementing the models within operational contexts; and the usability of these models once they are implemented. Each of these groups of criteria is discussed briefly in turn.

### *4.1. Credibility criteria*

This set of criteria deal with the basic confidence one has in a given model, as well as the suitability of the model as a policy analysis tool. Each criterion is briefly discussed.

#### *Theoretical soundness*

If a model is not theoretically sound, then one can have little confidence in its predictive capabilities and sensitivities. Aspects of theoretical soundness include

**Box 2**  
Model evaluation criteria

**Credibility**

- Theoretical soundness
- Policy sensitivity
- Precision (spatially, temporally)
- Validation

**Feasibility**

- Computational requirements
- Data requirements
- Technical support requirements
- Cost

**Usability**

- Ease of input preparation
- Model run time
- Output/presentation capabilities
- Portability/transferability
- Flexibility/adaptability

the following: the model captures key behavioral relationships and includes key actors, processes, etc.; the model is consistent with our current theoretical and empirical understanding of urban processes; the model is internally self-consistent; and the model makes use of statistically and logically valid methods and procedures.

*Policy sensitivity*

To be of practical use, the model must be capable of responding to the range of policy issues of interest. These can and should include land use policies (zoning, taxation, growth management policies, land servicing, etc.); transportation policies (transit, road, transport demand management, etc.); auto ownership related (vehicle technology options, taxation, etc.); and various combinations thereof. Put another way, the model should be able to analyze a wide range of infrastructure investment, operating, regulatory, and financial options, which target the demand and/or supply sides of the various processes/markets at work within the urban area.

*Spatial and temporal precision*

The model must be able to analyze the urban system and provide forecast outputs at sufficiently precise spatial and temporal scales to address policy questions in adequate detail. The definition of “sufficiently precise” of course varies from one

application to another. In some cases, very “broad brush” results at a very gross spatial scale over one or more very large time steps may well suffice. In others, much more spatial and/or temporal detail will be required. Ideally, the model should be able to “window in and out” to the level of precision required for a given analysis. In practice, however, such flexibility is difficult to achieve and may imply a very high cost in terms of program complexity, computational and data resources, etc.

### *Validation*

Rather than speak of model accuracy (which is a very difficult thing to determine in practical terms) it is perhaps more useful to speak of model validity. Validity is established in at least two ways. First, a model has face validity if its results generally are in agreement with best professional judgement and with empirically observed past and current trends. Second, a model can be historically validated in order to demonstrate its performance in replicating historical trends that lie outside the time period used in model calibration.

### *4.2. Feasibility criteria*

This set of criteria deal with the technical and financial feasibility of a modeling system. The three technical criteria have already been discussed at some length. Cost has been added as an explicit criterion, given its obvious importance in any decision process. Costs include both implementation and ongoing operating costs. In both cases, costs will exist for the model software system itself, computer hardware and ancillary software, data collection and maintenance, and the technical support staff (both in-house and external) required to implement and operate the model. For large-scale integrated urban models, both implementation and operating cost components can be significant and can represent a major constraint or even barrier to model adoption and usage.

### *4.3. Usability criteria*

Ease of use represents another important dimension for evaluating any model. Each of the “usability” criteria listed in Box 2 are briefly discussed below.

#### *Input data preparation*

Even conventional four-stage travel demand models typically have very onerous, time-consuming, and often error-prone input data preparation requirements,

which can significantly limit the number of alternatives investigated, the extent of sensitivity testing, etc., relative to what one would ideally like from a planning perspective. Integrated urban models can easily possess the same or even greater problems in this regard.

#### *Model run time*

It can be argued that within fairly broad limits, model run time is not overly critical in most applications: if it takes an overnight run to generate results that one might then spend a week or more analyzing, the actual run time is fairly unimportant. Very long run times, however, can be a major obstacle to the initial calibration of a model: if each model run with a given set of calibration parameters takes a night to execute, then it might literally take months to find the set of parameter values that “optimize” the performance of a model.

#### *Output/presentation capabilities*

Much more important is the question of result output and presentation capabilities. One runs a model in order to obtain useful outputs. The nature of these outputs, and the ease with which they can be stored, accessed, manipulated, and presented, are critical to the overall utility of the modeling system. Indeed, no matter how substantively credible the results might be, if they cannot be readily used to provide real insight into the problem being addressed, and if they do not easily translate into formats and messages that can be understood by non-technical decision-makers, then they are of little practical use. Models are ultimately decision support tools, and their effectiveness in this role depends in no small way on their output/presentation capabilities.

Given the power of ancillary software readily available today, much of this capability may reside outside of the modeling system *per se*. In such cases, the key concern is the ease of interface between the model and such software packages. In other cases, some or much of the post-run analysis and display capability may reside within the package itself, thereby minimizing the need to move data from place to place, as well as, perhaps, providing efficient, customized analysis and report generation capabilities.

Regardless of how the post-run analysis and display tasks are performed, however, the more fundamental concern is that the right information is computed and stored within the model for later retrieval and analysis. This, of course, returns one to fundamental questions of model design. In this case, however, the design is driven from the “bottom up” in terms of beginning by asking what information is needed from the model, rather than from the “top down” – as has implicitly been the case in much of the discussion to this point – in terms of what information the model can provide easily (given available theory, data, etc.).

### *Portability/transferability*

The greater the extent to which a model can be transferred from one application (urban area) to another, the more credible and useful it is, for at least three reasons. First, transferability implies generality, which, in turn, implies that the model must be theoretically sound since it seems to hold across a variety of applications. Second, development and implementation costs should be reduced, given that they are spread across several users. And third, shared models generate a community of users who can share experiences, help each other with common problems, and collectively contribute to model improvements over time.

### *Flexibility/adaptability*

Models are not static entities. Rather, they (should) grow and evolve as data, theory, computational capabilities, experience, financial constraints, etc., change over time. Similarly, application contexts are constantly evolving as new issues arise, new alternatives are suggested, decision-makers and their interests change, etc. Thus, models ideally should be flexible and adaptable so that they can evolve and change over time in response both to new challenges and new opportunities.

## **5. Summary**

This chapter has discussed general design requirements for the development of operational, policy-sensitive integrated land use/transport models. It has also presented a set of criteria that may be used to assess such models for possible use in an operational setting. Specific integrated models will deal with these design requirements in a variety of ways, ranging from ignoring one or more of these requirements altogether through to a fairly detailed treatment of most or all of the issues discussed in this chapter. It is fair to say that no currently operational model fully incorporates all aspects of what one might expect in an “ideal” model. It is also the case, however, that integrated models have been evolving rapidly in recent years toward this ultimate goal.

## **Acknowledgments**

The ideas presented in this chapter were first developed in collaboration with D. Kriger and J.D. Hunt as part of the US Transit Cooperative Research Program Project H-12, “Integrated Urban Models for Simulation of Transit and Land use Policies.”

## References

- Anas, A. (1995) "Capitalization of urban travel improvements into residential and commercial real estate: simulations with a unified model of housing, travel mode and shopping choices," *Journal of Regional Science*, 35:351–375.
- Badoe, D.A. and E.J. Miller (2000) "Transportation-land-use interactions: empirical findings and implications for modeling," *Transportation Research D*, 5:235–263.
- Ben-Akiva, M.E. (1974) "Structure of passenger travel demand models," *Transportation Research Record*, 526:26–41.
- de la Barra, T. (1989) *Integrated land use and transport modelling*. Cambridge: Cambridge University Press.
- Hensher, D.A. (2002a) "Integrated transport models for environmental assessment," in: D.A. Hensher and K.J. Button, eds, *Handbooks in transport*, Vol 4. *Handbook of transport and the environment*. Oxford: Pergamon.
- Hensher, D.A. (2002b) "A systematic assessment of the environmental impacts of transport policy," *Environmental and Resource Economics*, 22:1–2.
- Hunt J.D. and D.C. Simmonds (1993) "Theory and application of an integrated land-use and transport modelling framework," *Environment and Planning B*, 20:221–244.
- Knight, R.L. and L.L. Trygg (1977) "Evidence of land use impacts of rapid transit systems," *Transportation*, 6:231–247.
- Martinez, F. and P. Donoso (2001) "MUSSA: a land use equilibrium model with location externalities, planning regulations and pricing policies," in: *7th International Conference on Computers in Urban Planning and Urban Management (CUPUM 2001)*. Hawaii.
- Miller, E.J., D.S. Kriger and J.D. Hunt (1998) *Integrated urban models for simulation of transit and land-use policies*, Final Report, Transit Cooperative Research Project H-12. Toronto: University of Toronto Joint Program in Transportation ([www4.nas.edu/trb/crp.nsf](http://www4.nas.edu/trb/crp.nsf)).
- Southworth, F. (1995) *A technical review of urban land use – transportation models as tools for evaluating vehicle travel reduction strategies*, Report ORNL-6881. Oak Ridge: Oak Ridge National Laboratory.
- Waddell, P. (2000) "A behavioral simulation model for metropolitan policy analysis and planning: residential location and housing market components of UrbanSim," *Environment and Planning B*, 27:247–263.
- Wegener, M. (1995) "Current and future land use models," in: G.A. Shunk, P.L. Bass, C.A. Weatherby and L.J. Engelke, eds, *Travel Model Improvement Program Land Use Modeling Conference Proceedings*. Washington, DC: US Department of Transportation.
- Wegener, M. (2000) *IRPUD: the IRPUD model: overview*. Dortmund: Institut für Raumplanung.

*Chapter 11*

## LOWRY-TYPE LAND USE MODELS

ALAN J. HOROWITZ

*University of Wisconsin, Milwaukee, WI*

### 1. Introduction

Land use models allocate a wide variety of activities to zones of a region for the purposes of land use planning, regional economic planning, or travel forecasting. They contain residential location models and models of industrial and service location. Land use models attempt to:

- allocate residential, industrial, and service activities consistently with each other and consistently with transportation supply;
- resolve conflicts over available land for these activities.

There are many land use models, and they differ markedly in their internal workings. Historically, the most popular land use model was first sketched by Ira Lowry of the Rand Corporation in 1964. It has been enhanced and refined by many researchers, and it has become quite sophisticated in recent years. Notable examples of implementations of Lowry-type models include PLUM, TOMM, LILT, ITLUP, DRAM-EMPAL, and HLFM II+. In addition, the Lowry model is the conceptual ancestor of a broad class of urban activity models that combine input/output analysis with spatial choice theories, such as TRANUS and MEPLAN.

Land use forecasts can interact with travel forecasts. A travel-forecasting model provides many of the inputs to a land use model, such as zone-to-zone travel times. Furthermore, a travel-forecasting model can accept data about levels of urban activity from a land use model. It is possible for the two classes of model to be run together, iteratively.

The principle use of a Lowry-type model is to allocate a fixed amount of population and employment to zones of a region, given known locations of some of that employment and the transportation characteristics of the region.

The purpose of this chapter is to present an outline of the theoretical underpinning of Lowry-type models, to show how they can interact with conventional travel-forecasting models, and to discuss the strengths and weaknesses of the concept.

## 2. Land use model concept: urban form and land rents

### 2.1. *Urban form and land rents*

The process of urban development is highly complex, so available theories are only the barest approximation of reality. The earliest theories of urban form assumed that cities were monocentric (i.e. identical, regardless of the direction of travel from the central core) and that travel was equally fast between any two points. This “featureless” plain of development was first used by Johan von Thunen in 1826 to predict the location of crops around agricultural markets.

Von Thunen noticed that difficult-to-transport crops (e.g. vegetables) were grown close to markets on relatively highly priced land, while sturdier crops (e.g. grains) were grown at locations more distant from markets on lower-priced land. Von Thunen suggested that vegetable farmers were outbidding grain farmers for lands proximate to the market in order to reduce their transportation costs. Farmers in von Thunen’s time rented their land, so the amount paid to the landowner was referred to as a “land rent.” Today, land rent would be reflected in the selling price of the parcel. Von Thunen argued that there was a direct relationship between transportation costs and land rent.

William Alonso in 1960 showed that von Thunen’s theory of agricultural land also applied to urban residential locations. Alonso theorized that each business and each category of resident has a “bid rent” function, which is the relationship between distance from the city center and the amount of rent that can be paid. The various businesses and residences compete for land, with any given parcel of land going to the highest bidder. Bid rent functions vary according to the amount of land needed for a particular activity, the cost of travel to the city center, and individual taste.

Of course, real cities are not monocentric, but there is still a tendency for people to want to locate their businesses and residences near the city center. Although there are numerous counter-examples, land nearest the city center is usually more expensive than land farther away. As land becomes more expensive, it is used more intensively (unless it is prevented from more intensive development by land constraints, such as zoning or the longevity of existing development). The cost of land is affected by the cost of traveling a given distance, which is related to the quality and quantity of transportation supply.

A pure economic theory is much too difficult to apply in a typical multicentered city. Instead, researchers have constructed hybrid models that combine notions of statistics, probability theory, choice theory, and economic theory. These hybrid models behave in much the same manner as Alonso’s theory, but do so with a detailed zone system and transportation network. Hybrid models can capture many of the subtleties of urban activity location. The Lowry model is such a hybrid.

## 2.2. Agglomeration

Agglomeration is the tendency for businesses to locate close to one another – in business districts, on commercial strips, and in malls. There are many reasons for agglomeration, but the most important ones relate to locations of markets and sources of raw materials and required services:

- Businesses operate more efficiently if they are near required services. For example, central business districts are highly desirable locations for offices because of their concentrations of lawyers, accountants, etc.
- Retail establishments tend to locate near the center of their market areas. Competing establishments do not necessarily divide their shared market evenly. Instead, there is a tendency for competing establishments to locate near one another.
- Retail establishments may try to locate near places where there are already customers. By doing so, retail establishments can capture a share of customers from competing establishments, lure customers from complementary establishments, and try to increase the total number of customers who would visit that general location.

Agglomeration is one of the strongest factors in determining how urban land is used, and agglomeration is strongly affected by transportation supply. Unfortunately, it is often difficult to predict where new activity centers will arise or to determine the eventual size of the center.

## 3. Residential location models

### 3.1. Basic form

The simplest residential location model is a form of the gravity model – the same type of model used to perform trip distribution in travel-forecasting models. In this case, however, trips are produced at the workplace and attracted to home. This definition of productions and attractions is exactly opposite of the one used for conventional travel forecasting. In essence, the model assumes that workplaces have fixed locations, but residences can move around. Hence, the purpose is work-based home (wbh). Thus,

$$T_{ij(\text{wbh})} = \frac{e_i w_j f(t_{ij})}{\sum_j w_j f(t_{ij})}, \quad (1)$$

where  $w_j$  is the residential attractiveness of zone  $j$  (not the number of trip attractions),  $e_i$  is the employment in zone  $i$ ,  $t_{ij}$  is the disutility of travel from zone  $i$

to zone  $j$ , usually expressed in units of time, and  $f(t_{ij})$  is a deterrence function value (similar to a friction factor) for a trip from zone  $i$  to zone  $j$ .

The deterrence function,  $f(t_{ij})$ , is a strictly declining function of disutility, and it accounts for people's reluctance to travel long distances to work. The attractiveness of a zone,  $w_j$ , is a measure of the ability of the zone to draw trips, and is closely related to the amount of land available for or already in residential development. Attractive zones are not necessarily pretty. More will be said about attractiveness later in this section.

The disutility of travel is a concept similar to path impedance. It can contain terms for:

- travel time between zones;
- en route travel costs, including tolls and fuel;
- terminal times and charges.

Equation (2) embodies three principles:

- workers tend to locate their residences close to their workplaces, all other things equal;
- zones with relatively large amounts of residential land tend to attract a relatively large proportion of workers' residences, all other things equal;
- the measure of closeness, disutility, includes both the quantity and quality of the transportation system.

It is inappropriate to say that workers single-mindedly try to minimize their travel disutility (time, etc.) to work. Rather, travel disutility is one of many factors that workers consider when choosing residential locations. Workers also prefer good schools, familiar neighborhoods, and decent housing. The strength of workers' desires to live near their job sites is embodied in the deterrence function,  $f(t_{ij})$ , which is always calibrated to local data.

The most popular deterrence function is of the form

$$f(t_{ij}) = \exp(-\beta t_{ij}), \quad (2)$$

although other forms have been used occasionally. This deterrence function was dubbed an "entropy-maximizing model" by Wilson (1967), who was able to derive it from principles of probability theory. Alternatively, DRAM, a residential location model discussed later, uses a slightly more complex deterrence function of the form

$$f(t_{ij}) = t_{ij}^\alpha \exp(-\beta t_{ij}), \quad (3)$$

although the authors of DRAM have reported that either the exponent  $\alpha$  or  $\beta$  can be set to zero without greatly affecting the results. With  $\alpha$  set to zero, eq. (3) is identical to eq. (2).

The portion of the residential location model without the employment in zone  $i$  is the probability that a worker from zone  $i$  will locate his/her residence in zone  $j$ . This may be written as

$$p_{ij} = \frac{w_i f(t_{ij})}{\sum_j w_j f(t_{ij})}. \quad (4)$$

Equation (4) closely resembles models used for mode choice. If the deterrence function is allowed to be eq. (2), eq. (4) is essentially identical to a spatial choice model that is based on a multinomial logit function.

Consequently, the number of work-based home trips between zones may be found as

$$T_{ij(\text{wbh})} = e_i p_{ij}. \quad (5)$$

The number of workers residing in zone  $j$  is simply

$$e_j^r = \sum_i e_i p_{ij}, \quad (6)$$

provided all employment sites have been included in the set of employment zones,  $i$ . The population of zone  $j$  is

$$p_j = q_j e_j^r, \quad (7)$$

where  $q_j$  is the number of people for each employee living in zone  $j$ . This equation for population contains a subtle error that is almost always ignored – the total population of the region can depend upon transportation supply. For example, a redistribution of some workers from a zone with a small  $q_j$  to a zone with a large  $q_j$  can cause an overall increase in population. The error is tolerated because household size is more strongly related to the zone of residence than to the zone of employment.

### 3.2. Population segmentation

To the extent that suitable employment data is available, it is possible to segment workers within the residential location model. Workers could be segmented by race or ethnic group or income, depending upon the goals of the travel forecast. The residential location model would be run separately for each segment, and the results from the various segments would be added together. Planners should proceed with caution when segmenting a residential location model, so as to not inadvertently endorse current patterns of racial or ethnic segregation. Segmented models have a tendency to preserve existing patterns of segregation in the forecast.

### 3.3. Measures of attractiveness

The strongest single measure of residential attraction is the zone's residential developable area – the amount of land available for residential development, including land already devoted to residential use. The developable area of a zone is most easily calculated by subtracting unsuitable land parcels from the gross area of the zone. Unsuitable lands include parcels already devoted to industrial or commercial uses, environmentally sensitive areas, institutional lands, parks, preserved lands, quarries, and lands precluding residential construction (unsuitable soils, high water table, steep slopes, water coverage, or exposure to high levels of pollution).

DRAM (the disaggregate residential allocation model) uses an expanded measure of attractiveness for any zone,  $j$ :

$$w_j = L_j^q X_j^r R_j^s N_{j1}^t N_{j2}^u N_{j3}^v N_{j4}^w, \quad (8)$$

where  $L_j^q$  is the vacant developable area for the zone,  $X_j^r$  is one plus the fraction of developed land area that is already developed for the zone,  $R_j^s$  is the residential land in the zone, and  $N_{j2}^u, N_{j3}^v, \dots$  are one plus the fraction of employed residents in the 1, 2, ... income quartile.

The four "N" terms capture the joint effects of economic prosperity and residential density. The parameters  $q, r, s, t, u, v$ , and  $w$  are empirically derived.

DRAM has undergone extensive testing by Putman and his co-workers at the University of Pennsylvania (Putman, 1983). DRAM has elements of income segmentation, but does not require itself to be executed separately for each segment. The parameters will vary from city to city and from zone system to zone system, so they will require special calibration. However, a good starting point for calibration (and for understanding how the model works) would be about  $q = 0$ ,  $r = 1.5$ ,  $s = 0.9$ ,  $t = 4.7$ ,  $u = 1.6$ ,  $v = -3.0$ , and  $w = -0.8$ . In order to properly use DRAM, good knowledge of the state of development in the forecast year is required, so it is best suited for short- or medium-term forecasting.

### 3.4. Land constraints

The residential location models, as described, are unconstrained. That is, they could allocate more population to a zone than would be allowed under current zoning. If effective land constraints are to be included in the forecast, they must be handled by:

- making arbitrary reductions in zonal attractiveness;
- increasing travel times to the zone (including intrazonal time) for the work-based home trip purpose.

Reducing zonal attractiveness is the easier of the two methods because there is an almost exact one-to-one sensitivity between attractiveness and population.

### *3.5. The exogenous workplace*

People who build simulation models divide all variables into two groups: endogenous and exogenous. The values of endogenous variables are determined by the model, and the values of exogenous variables are set by the user. In a residential location model, workplaces are exogenous while residential locations are endogenous. Such a model follows traditional urban theory by assuming that residences are located relative to work sites.

Some observers of urban development have noticed a new trend in workplace location. Many employers are seeking a site that is located near a suitable pool of labor. No longer can workplaces be considered strictly exogenous, because some types of employment might shift with the relocation of residences. A consensus has not yet been reached as to the significance of this trend for land use forecasting.

### *3.6. Multimodal applications*

When a pair of zones is well served by more than one mode, it is possible to include all modes in the disutility term. Practice has been to compute a composite disutility with a “log sum” expression that is similar to those used in nested logit models of mode split. The effect of adding transit service is to reduce disutilities found from a highway network and to cause a redistribution of populations and services toward areas served by transit.

## **4. Overview of the Lowry model**

A Lowry-type model essentially adds just one new feature to a residential location model – service location. Within a Lowry-type model, services are defined as those employers who derive their income from within the region and who are sensitive to the locations of their customers. Services are further subdivided into two classes:

- those that serve people and tend to locate proximate to concentrations of population;
- those that serve businesses and tend to locate proximate to concentrations of employees at their workplace.

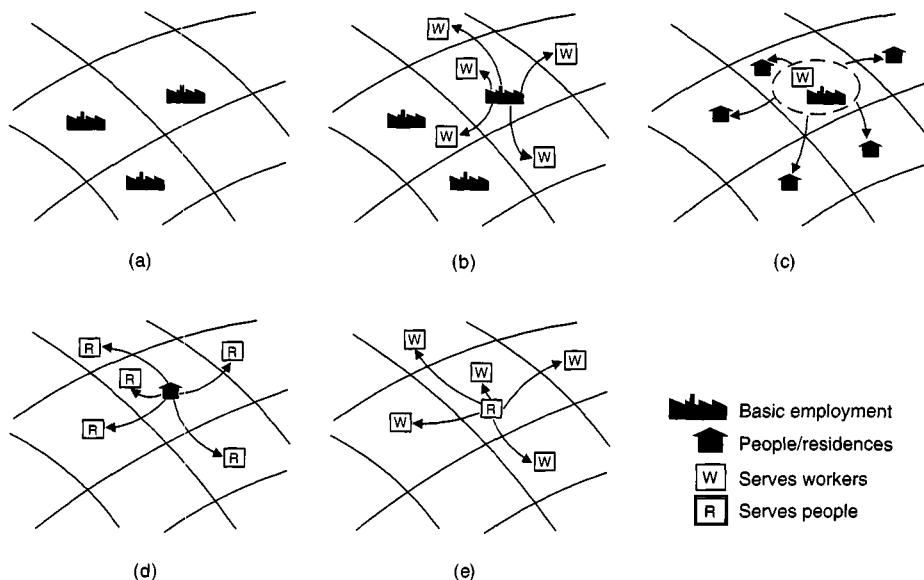


Figure 1. The Lowry model concept.

Services are allocated to zones in much the same way as residences are allocated to zones, considering both service attractiveness and the disutility of travel.

A Lowry-type model cannot allocate “basic” industries (businesses that derive their income from outside the region), so their locations must be provided as input.

Lowry-type models become computationally awkward because:

- services must themselves be served;
- services have employees needing residences.

Consequently, Lowry-type models simultaneously solve for the number of people and the number of service employees in every zone. Such a solution requires a large amount of computation, especially if the models must also resolve conflicts over land or satisfy hard constraints on population or on service employment.

Figure 1 illustrates the conceptual operation of a Lowry-type model. Figure 1a shows the given allocation of basic employees to zones. Figure 1b shows the allocation of services that serve basic industries. Figure 1c shows the allocation of population to zones and Figure 1d shows the allocation of services that serve residences. Figure 1e shows the allocation of services that serve other services. As an expedience, most Lowry-type models do not make a distinction between the types of services that serve basic industries and the types of services that serve other services, although doing so would not be especially difficult.

Once population and services have been allocated, it is possible to perform a traffic forecast in the usual way. The traffic forecast may reveal unanticipated congestion effects, so the land use model may have to be executed again.

#### *4.1. Typical data requirements*

Since land use forecasts are typically performed on a coarse zone system, the total amount of zonal data is somewhat less than for a standard travel forecast of the same city. However, a land use forecast requires the same data about the highway network as a standard travel forecast: a topological description of the network (links and nodes), link travel times, capacities, turn penalties, etc. At the very least, zonal data consist of:

- residential attractiveness;
- service attractiveness;
- basic industry employment at the zone of employment;
- the ratio of population to employees at the zone of residence.

Service attractiveness is closely related to zone size. Depending upon the objectives of the forecast, additional zonal data may be necessary:

- hard constraints on population or service employment;
- the ratio of service employees that serve residences to population;
- the ratio of service employees that serve businesses to employment;
- the amount of land required by each service employee.

Like a standard travel forecast, a land use forecast needs a certain amount of careful calibration. The calibration step will indicate which of the zonal data need to be included in the forecast and which zonal data need to be adjusted for intangible, but important, factors.

Additional data requirements often include:

- the monetary cost of automobile travel;
- the value of time;
- average trip lengths for work-to-home, home-to-service, and work-to-service trips;
- locations of services that are unlikely to move.

#### *4.2. Anticipated results and cautions*

A land use forecast will yield the following results:

- the population, service employment, and total employment of each zone;

- modified values for residential and service attractiveness that account for constraints.

The land use model will also provide summary statistics and diagnostic information, which may be helpful in further calibration and in finding problems.

When running a land use model it is important to recognize its limitations, and the results must be properly interpreted to avoid misrepresentations:

- The model forecasts equilibrium conditions, so the forecast year is indefinite. It is the planner's responsibility to determine the amount of change that would take place by any given point of time.
- Activities are spread evenly across zones. The model cannot reveal where concentrations of services will occur. A separate procedure would be necessary to reallocate activities to smaller zones.

#### *4.3. Calibration issues*

Like standard travel-forecasting models, land use models require a significant amount of calibration. Essential ingredients in a good calibration are data about existing land use, demographics, socio-economic characteristics, and travel conditions.

#### *4.4. Equilibrium conditions*

Lowry-type models predict equilibrium conditions for development, so they should ideally be calibrated from data representing equilibrium conditions. Unfortunately, cities are never in equilibrium, always changing in response to economic, social, and transportation trends. It may take several decades for a city to fully adjust to a major new transportation facility, such as a new freeway or a new fixed-rail transit line. Great care must be exercised to adjust all base year data for consistency, so that the base year forecast does not closely predict existing land use but instead predicts the full land use impact of the existing transportation system. The model is working properly if it over-predicts population in zones that are currently growing and if it under-predicts population in zones that are currently declining.

#### *4.5. Deterrence function parameters*

Most planning studies can get by with a deterrence function having a single parameter. It has been determined that a good value for that parameter can be

found by getting the model to replicate the actual average trip length for the corresponding trip purpose. Finding the single parameter of an exponential deterrence function,  $\beta$ , usually requires only a few runs of the model, because the average trip length is almost equal to the reciprocal of  $\beta$ . Deterrence functions with multiple parameters require advanced statistical techniques.

#### *4.6. Disutility and the value of time*

The most important parameter of the disutility function is the value of time, which is a means of converting monetary costs to time units. The value of time is almost impossible to determine by brute-force calibration of location models. The best sources for values of time are mode split studies. If a logit (or similar) model has been statistically calibrated, it is possible to compare the parameters for in-vehicle time and money. The ratio of these two parameters is the implied value of time for travelers considering transit.

In the absence of a calibrated mode split model, values of time must be adopted from other cities. Researchers have found that the value of time on work trips is between 25% and 50% of the prevailing wage rate. Old values of time should be properly inflated.

#### *4.7. Definition of basic employment*

In a Lowry-type model it is possible to vary the definition of basic employment to achieve a better match to existing conditions. Services that are unresponsive to their markets or are strongly influenced by agglomeration can be lumped with basic industries. These services can be directly allocated to zones by the planner. Hospital complexes, regional shopping malls, and central business districts can be represented in this fashion. On the other hand, basic employers that are assumed to move relative to their labor pools can be considered together with service employment. Care should be taken to set the various model parameters according to the adopted definition of “basic” and “service” employment.

### **5. Derivation of the Lowry–Garin model**

The first major refinement of the Lowry model was suggested by Robert Garin in 1966. Garin replaced some of Lowry's iterative procedures with matrix equations. The following description of the Lowry–Garin model is confined to those features commonly found in Lowry-type models.

The Garin version of the Lowry model is a series of matrix equations that forecasts the distribution of population and employment in an urban area. The Lowry-Garin model is most easily derived by constructing an employment conservation equation. Let  $\mathbf{E}$  be a vector of total employment (each element,  $e_i$ , of  $\mathbf{E}$  being the total employment of the  $i$ th zone),  $\mathbf{E}_B$  be a vector of basic employment,  $\mathbf{E}_R$  be a vector of service employment required by residences, and  $\mathbf{E}_W$  be a vector of service employment required by workers (i.e. businesses). Total employment is thus the sum of its three components:

$$\mathbf{E} = \mathbf{E}_B + \mathbf{E}_R + \mathbf{E}_W \quad (9)$$

Each of the three vectors on the right-hand side of eq. (9) represent the spatial distribution of a sector of employment in the urban area. Basic employment,  $\mathbf{E}_B$ , is the only explicit exogenous variable in the Lowry-Garin model. Employment serving residences,  $\mathbf{E}_R$ , and employment serving workers,  $\mathbf{E}_W$ , are dependent upon trip-making patterns, the transportation system, and existing land use. Although Lowry was concerned with only three employment sectors, eq. (9) is easily extensible to many more sectors.

Employment in industries that serve workers,  $\mathbf{E}_W$ , is calculated by distributing service employees around all employment locations as given by the vector  $\mathbf{E}$ . Define  $h_{ij}$  as the conditional probability that an employee in zone  $j$  is served by another employee in zone  $i$ , the resultant matrix of conditional probabilities as  $\mathbf{H}$ , and the number of service employees required for each employee, averaged across the whole urban area, as  $f$ . Then

$$\mathbf{E}_W = f\mathbf{HE}. \quad (10)$$

Some models permit specifying a different  $f$  at each zone.

A similar relation can be constructed for employees serving the entire population. Define  $b_{ij}$  as the conditional probability that an individual who lives in  $j$  is served by an employee in  $i$ , and the conditional probability matrix as  $\mathbf{B}$ . Also, define  $g$  as the number of employees that serve each individual, averaged across the whole urban area. Then, as in eq. (10),

$$\mathbf{E}_R = g\mathbf{BP}, \quad (11)$$

where  $\mathbf{P}$  is the population vector, containing elements,  $p_i$ , each of which is the population in zone  $i$ . The variable  $g$  could have been set separately for each zone, but was not. The population distribution is computed from total employment. Define  $a_{ij}$  as the conditional probability that an individual working in  $j$  lives in  $i$ . Let  $\mathbf{A}$  be the matrix of these conditional probabilities. Also, define  $q_i$  as the ratio of population to employees in residential zone  $i$ . Furthermore, let

$$\mathbf{Q} = [\delta_{ij}q_i], \quad (12)$$

where  $\delta_{ij}$  is the Kronecker delta. Note that  $\mathbf{Q}$  is a diagonal matrix. Populations of all the zones are found from

$$\mathbf{P} = \mathbf{QAE}. \quad (13)$$

Consequently, from eqs (11) and (13),

$$\mathbf{E}_R = g\mathbf{BQAE}. \quad (14)$$

Substituting eqs (10), (11), and (14) into eq. (9) reduces the employment conservation equation to one with terms for only total employment,  $E$ , and, basic employment  $\mathbf{E}_B$ , i.e.

$$\mathbf{E} = \mathbf{E}_B + g\mathbf{BQAE} + f\mathbf{HE}. \quad (15)$$

Equation (15) can be solved for the spatial distribution of total employment,  $\mathbf{E}$ , in terms of basic employment. Specifically,

$$\mathbf{E} = (\mathbf{I} - g\mathbf{BAQ} - f\mathbf{H})^{-1}\mathbf{E}_B. \quad (16)$$

The spatial distribution of population can be computed from eq. (13), and the spatial distributions of employment in the two service sectors,  $\mathbf{E}_R$  and  $\mathbf{E}_W$ , are directly computed from eqs (10) and (14).

The three conditional probability matrices ( $\mathbf{A}$ ,  $\mathbf{B}$ , and  $\mathbf{H}$ ) are computed from singly constrained trip distribution equations with any deterrence function. For example,  $\mathbf{A}$  can be found from

$$a_{ij} = \frac{w_i \exp(-\beta t_{ij})}{\sum_i w_i \exp(-\beta t_{ij})}, \quad (17)$$

where  $t_{ij}$  is the disutility of travel between zones  $i$  and  $j$ ,  $w_i$  is the attractiveness for residential zone  $i$ , and  $\beta$  is a calibration parameter. Disutility of travel has terms for travel time and travel cost.

### 5.1. Adjustments to residential attractiveness

Land area does not formally appear in either the Lowry–Garin model or in the trip distribution equations. Nonetheless, land area can be introduced into the model by making residential attractiveness ( $w_i$  in eq. (17)) equal to residential developable area and by making service attractiveness equal to service developable area. The same parcel of land may be included in both measures of attractiveness. The trip distribution equations will assign activities to zones roughly in proportion to these developable areas. If a zone is almost fully occupied by service activities, then only a few people should be able to live there. The model can be instructed to reduce residential trip attractiveness in response to large allocations of service employees to a zone.

In addition, some models can be forced to allocate a specific population to any given zone. This constraint is satisfied by making adjustments to residential

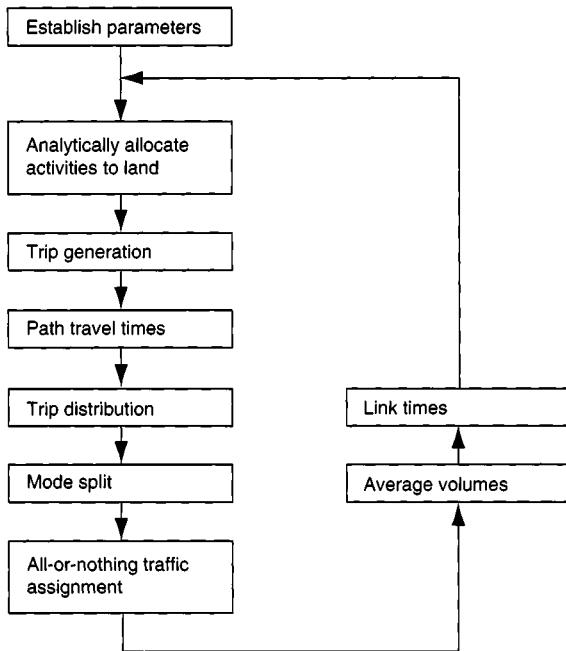


Figure 2. Land use forecasting with travel forecasting.

attractiveness in the trip distribution equations. It is not easy to involve land price in the calculations.

### 5.2. *Adjustments to service attractiveness*

Some models can be instructed to allocate a specific number of service employees to any given zone. This constraint is satisfied by making adjustments to service attractiveness within the trip distribution equations.

## 6. Iterating a land use model with a travel-forecasting model

A Lowry-type model can be run stand-alone or within a travel-forecasting framework. Integrated models have been growing in popularity since the first reported application of ITLUP in the 1970s, particularly outside of the USA.

Figure 2 shows one method of how a land use model might be iterated with a travel-forecasting model. The land use model shows up as an activity allocation step in the flow diagram.

The feedback loop assures that the land use forecast is properly sensitive to levels of congestion on the network. A “volume-averaging” step is necessary to ensure that the model converges to an equilibrium solution. Volume averaging can be performed in a number of ways, but the most practical method is called the method of successive averages (MSA). MSA takes an unweighted average of all of the all-or-nothing traffic assignments. MSA seems to work impeccably on problems where the trip table is allowed to vary with volume, including land use forecasting.

It should be apparent from the structure of an integrated Lowry-type model that it is best suited for forecasting the gross effects on land development patterns from changes in transportation infrastructure or policies. For example, such a model could reveal the degree of population decentralization that might occur when a freeway is widened, and the effect of this decentralization on levels of congestion elsewhere on the highway system.

## 7. Critique

The fundamental purpose of a land use model is not to predict the future but to aid long-range transportation, land use and environmental decision-making. Thus, a model need not be a perfect representation of reality, so long as it helps reach the correct decision. Thus, it is important that analysts who prepare land use forecasts have a good understanding of the strengths and weaknesses of their models. A few aspects of Lowry-type models that could make them less than ideal in some planning situations are:

- *Worker location and mobility.* Lowry-type models assume a great deal of temporal stability in workers job location and assume that employment location depends only on prior location decisions of basic employers and the markets of service employers. The location of residents or other non-market factors cannot directly influence job location.
- *One-to-one relationship between worker location and residential location.* Residential location models were initially developed during a time when there was typically one worker in a household. These formulations can only approximate the effect of multiple-worker households on residential location decision.
- *Weak relationship to historical development patterns.* Except in a few places where the analyst can influence input data, Lowry-type models do not use existing or earlier-than-forecast-year land use patterns. The cityscape is cleared and rebuilt each time the model is run.
- *No land prices.* While it is obvious that land prices have a very strong influence on urban development patterns, Lowry-type models do not have

mechanisms for determining land prices or of readily using land price information in location decisions. Lowry-type models also do not recognize building age, condition, or other factors between now and the forecast year that would affect redevelopment decisions.

- *Instantaneous effect of transportation infrastructure or policy changes.* Because Lowry-type models react immediately to any change in input data, normal time lags between a change and its effects are not apparent.
- *Lack of certain agglomeration effects.* Lowry-type models do not have formal mechanisms for handling certain types of agglomeration, such as the formation of retail districts and malls.
- *Intangible effects on zonal attractiveness.* Land area by itself may not be a suitable measure of zonal attractiveness. Intangible attributes of land parcels (such as natural beauty, prestige, proximity to shopping, quality of schools, density of existing development, and zoning) are of lower importance than land area when determining residential attractiveness, but might be important to determining the distribution of population across a region. Including such intangible effects in a Lowry-type models can be difficult.

Over the years, planners have found ways of working around these limitations, although many of the methods are *ad hoc*.

## 8. Closure

The early theoretical successes of the 1960s in the development of Lowry-type models were quickly followed by disappointment in the results of applications of land use models in general in the 1970s. Land use models had been oversold. New applications of land use models almost ceased in the USA, but efforts to implement land use models and further develop the concept continued in other countries. The last decade has seen a resurgence of interest in integrated land use models, including those descended from Lowry's original model, as traffic congestion and environmental concerns have grown.

## References

- Alonso, W. (1960) "A theory of the urban land market," *Papers and Proceedings of the Regional Science Association*, 6:149–157.  
Garin, R.A. (1966) "A matrix formulation of the Lowry model for intrametropolitan activity allocation," *Journal of the American Institute of Planners*, 32:361–364.  
Lowry, I.S. (1964) *A model of metropolis*, Report RM 4125-RC. Santa Monica: Rand.

- Putman, S.H. (1983) *Integrated urban models: policy analysis of transportation and land use*. London: Pion.
- Wilson, A.G. (19??) "A statistical theory of spatial distribution models," *Transportation Research*, 1:253-269

*Chapter 12*

## ECONOMETRIC MODELS OF LAND USE AND TRANSPORTATION

MARCIAL ECHEIQUE

*University of Cambridge*

### 1. Introduction

Computable models of large-scale spatial systems emerged from the pioneering work of Lowry (1964). He developed one of the earliest land use models that was able to forecast the location of population and employment for the city of Pittsburgh, USA, that had been subdivided into a large number of zones. The ability to predict the land uses at small scale was helpful for practical transport studies. Transport models had been in existence for more than a decade before the work of Lowry, but required strong assumptions about the future location of households and jobs to estimate travel patterns between the zones of a city. Land use models provided the required inputs to transport models, as well as estimating the likely impact of changes in the provision of transport facilities on the location of land use. The combination of land use models with transport models generated a flurry of applications of integrated land use/transport interaction models (Batty, 1976) for practical planning studies.

Early models were theoretically weak and required vast data sets for estimating by statistical means the basic relationships between the modeled factors. Difficulties with computer power, data availability, and the unreliability of the predictions led to disillusionment with large-scale models (Lee, 1973), and the abandonment of many. However, transport models continued to be highly popular around the world, despite the lack of reliable forecasts of land use changes needed for the models.

Since that time there has been considerable development of the theoretical basis for integrated models coupled with improvements in computer power and development of data sets. All of these have resurrected the interest and usefulness of large-scale spatial models (Webster et al., 1988; Wegener, 1994).

## 2. Theoretical foundations

The early models used analogies to physics to describe the behavior of firms and households in space. Spatial interaction such as migration of population, residential location, services location, or the distribution of trips from origin zones to destination zones were based on Newton's gravity law. Later, Wilson (1970) derived spatial interaction models from entropy maximization procedures. But the work of McFadden (see Domencich and McFadden, 1975) firmly anchored spatial interaction models into micro-economic theory.

The basis of the McFadden approach is the so-called random utility theory. The models state that individuals making a discrete choice (e.g. where to live, what mode of transport to use, where to shop, etc.) will choose the option that maximizes their utility, subject to random variations. The variation around the maximum utility follows a Weibull distribution. The derivation leads to the multinomial logit model, where the probability of choosing an option is proportional to the exponential of the utility given by the option, i.e.

$$p^o \propto \exp(\lambda u^o), \quad (1)$$

where  $p^o$  is the probability of choosing option o,  $\lambda$  is the parameter to be estimated and relates to the randomness of the distribution, and  $u^o$  is the utility of option o.

The randomness of the distribution can be interpreted in different ways, some of which are as follows:

- aggregation errors in the descriptions of the options within the model, which leads to random variation;
- errors by the modeler of the measurement of the utility that affects the behavior of individuals;
- individuals are not “optimizers” but “satisfiers” that lead to random fluctuation of the outcomes.

Probably a combination of all of the factors described above influences the randomness of the probability of the outcomes.

Williams (1977) demonstrated that models derived from entropy maximization à la Wilson are the same as those derived from utility maximization à la McFadden. He also demonstrated that the aggregate utility estimated from a set of discrete choices is equal to the log sum of individual choices, i.e.

$$\tilde{u} = \frac{1}{\lambda} \log \sum_o \exp(\lambda u^o), \quad (2)$$

where  $\tilde{u}$  is the aggregate utility,  $u^o$  is the utility of choice o, and  $\lambda$  is a parameter related to the randomness of the choice. This result is significant for spatial models where a consumer has a number of alternatives to choose from.

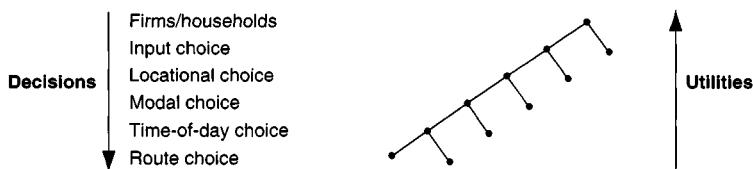


Figure 1. Hierarchically nested multinomial models.

The development of McFadden and Williams gives the theoretical basis for modeling the spatial economy. The models then reflect the behavior of firms and households choosing from a set of discrete options at different levels of a hierarchy of decisions (Figure 1). At the top level of the hierarchy, the decision is how much to consume of goods, services, land, etc. At the next level, the decision is where to obtain the inputs from (locational choice). At the next level, the decision is how to travel to obtain the inputs (modal choice). The next level is when to travel (time of day choice), and finally which route to take (route choice).

At each level of the hierarchy of decisions the choice is based on the utility given by each option. The utility reflects the trade prices, which includes the production price plus the transport price from where the production is to where it is consumed. Prices can also take into account other non-monetary costs. The utility at each level of the decision tree is the aggregate utility given by the choices taken at a lower level of the hierarchy. Williams' (1977) derivation of the aggregate utility gives the basis for obtaining the prices that determine the utility of each choice.

The model above estimates the demand for inputs assuming producers will provide them at a minimum price. In reality, there are constraints to production such as land and transport infrastructures that impose limits on how much can be used or, sometimes, because there are restrictions imposed by authorities (e.g. land use regulations). The introduction of physical or regulatory constraints in the modal frameworks (Echenique, 1994) gives rise to "pure rents" in the Marshallian sense. The rents reflect the congestion in the property and transport markets, and provide the mechanisms to bring the demand in line with the available supply (market clearing). The introduction of supply constraints within demand models gives rise to general equilibrium models of the spatial system.

Physical constraints can be overcome by changing the supply through time. Models of real estate or property development can be introduced into models, giving temporal dynamics to the general equilibrium model. More often than not the supply constraints of building and transport are determined exogenously, reflecting policy decisions whose impacts are to be assessed through the models.

One form of structuring the model of the spatial economy is through the framework of input-output analysis (Leontief, 1951). The framework is a true

system model that relates in a consistent form to all the elements of the spatial economy that are to be modeled. The framework has been used in regional economics for a long time, but with fixed demand and trade coefficients. The ability to combine input-output analysis with discrete choice theory gives rise to consistent computable general equilibrium models of the spatial system. Theoretical general equilibrium models that try to estimate prices and quantities simultaneously have a long history. But only with the advent of computers have the models become operational. There are theoretical examples of such models (Bröcker, 1995), but the model explained in the next section has been in use, albeit in partial form, since 1977 for practical planning studies (Echenique, 1994).

### 3. A general model of trade and location

#### 3.1. Functional relationships

Consider an economy made up of firms producing goods and services and households that consume these goods and services and sell labor. For practical reasons, such as data availability, the firms and households are aggregated into economic sectors and socio-economic groups, respectively. The sectors and groups are described as  $f$  factors of production ( $1, 2, 3, \dots, m, n, \dots, f$ )<sup>a</sup>. The level of factor  $m$  at the place of production is described as  $X^m$ , usually measured by the monetary or physical value of production. Physical values of production such as tonnage, employment or household numbers are valid. The transaction of a factor  $m$  from production to consumption by factor  $n$  is described by  $T^{mn}$ , normally measured by monetary flow, but could be measured by physical units (such as tonnage or person travel), i.e.

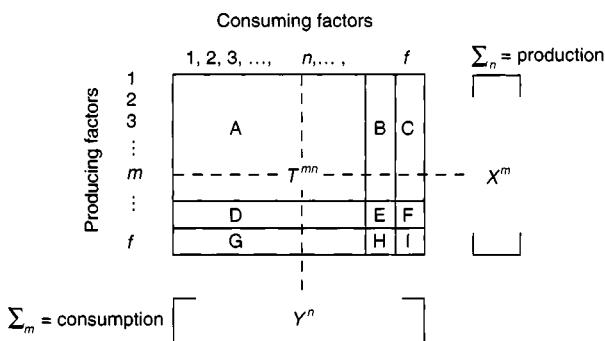
$$X^m = \sum_n T^{mn}, \quad (3)$$

$$Y^n = \sum_m T^{mn}, \quad (4)$$

where  $X^m$  is the total production of factor  $m$ ,  $Y^n$  is the total consumption by factor  $n$ , and  $T^{mn}$  is the transaction of factor  $m$  to be consumed or used by factor  $n$ .

It is useful for modeling purposes to partition the matrix  $T^{mn}$  in different sections, as described in Figure 2:

<sup>a</sup>The primary factors of production in economics usually include labor, land and capital. In this model, factors include any goods or service produced, labor, land and buildings, etc., that is to say, any element modeled in the system.

Figure 2. Transaction matrix  $T$ .

- Section A of the matrix  $T^{mn}$  represents the transactions between factors. This area is normally included in standard input-output models (Leontief, 1951). It represents the sales from sector  $m$  to sector  $n$ .
- Section B of the matrix  $T^{mn}$  represents the transactions between factor  $m$  and the household group  $n$ , in other words the consumption by the households of products or services  $m$ .
- Section C of the matrix  $T^{mn}$  represents the transactions between factors  $m$  to be exported to outside the area in consideration. Normally, both sections B and C are considered the final demand in standard input-output models that also includes investments and government consumption. It is described as the exogenous sector, that is to say, it is determined outside the model.
- Section D represents the sale of labor or other income received by socio-economic groups  $m$  from the factor  $n$  (e.g. dividends).
- Section E represents the sale of labor from socio-economic groups  $m$  to households in socio-economic groups  $n$  (e.g. domestic labor).
- Section F represents the sale of labor or other income received from the exogenous factor, such as pensions and other payments from government, etc.
- Section G represents the imports from outside the area and payments to the exogenous factor such as taxes to the government. In this sector, rental of property or land is sometimes included.
- Section H represents the payments by the households factor such as taxes, rental, etc.
- Section I represents payments by the exogenous factor to itself, such as imports for the government or for investments.

The following considerations are useful for adapting standard input-output models to this framework:

- Ideally the input–output definition should be related to activities producing goods and services that are tradable and thus easily identified for the transaction matrix.
- Consumption of goods and services by the household factor should not be ascribed directly to the households, but to commerce and other services providers. The household sector buys the goods and services normally through intermediaries in shops and other commercial establishments.
- Include the household factor into the endogenous transaction matrix that is to be modeled. The final demand thus contains only export outside the area, government and investment expenditure.
- Care should be taken to treat the transport factor and property factor in a manner consistent with the treatment in the rest of the model.

Given the transaction matrix  $T^{mn}$  that describes the trade between the factors, it is possible to estimate the functional relationships between the factors by establishing a matrix of coefficients  $a^{mn}$  representing the necessary input of activity  $m$  to produce one unit of output  $n$ . In a standard input–output model these coefficients are called technical coefficients, i.e.

$$a^{mn} = T^{mn}/Y^n, \quad (5)$$

where  $a^{mn}$  is the functional relationship between factors  $m$  and  $n$ , representing the necessary input of factor  $m$  to produce one unit of output  $n$ ,  $T^{mn}$  is the transactions between  $m$  and  $n$ , and  $Y^n$  is the total consumption or payments by factor  $n$  that is equal to total production  $X^m$  when  $n = m$ .

In standard input–output models the  $a^{mn}$  coefficients are considered fixed through time. This simplifying assumption is not necessary in general models, and it is possible to adjust the coefficients through time, depending on the relative prices of inputs  $m$ . Although there are minimum technical requirements of input  $m$  to produce output  $n$ , such as iron ore to produce steel, the quantities can be varied according to the prices of the input. Sometimes they can be substituted by other inputs such as recycled metal in the case of steel, i.e.

$$a^{mn} = f(p^m), \quad (6)$$

where  $a^{mn}$  is the functional relationship between factors  $m$  and  $n$  (input–output or demand coefficient) and  $f(p^m)$  is the function of price of factor  $m$ , which will be discussed in Section 3.4.

### 3.2. Spatial relationships

Consider an area divided into  $z$  zones ( $1, 2, 3, \dots, i, j, \dots, z$ ). At least one represents the rest of the world where exports go to and imports come from. The transaction matrix considered above can now be subdivided into spatial units

representing the trade of factors between zones. Now the production of factor  $m$  in zone  $i$  is described as  $X_i^m$ . Similarly, the transaction or trade (as it is now spatial) of factor  $m$  to be consumed or used by factor  $n$  from the zone of production  $i$  to the zone of consumption  $j$  is described by  $T_{ij}^{mn}$ :

$$X_i^m = \sum_j \sum_n T_{ij}^{mn}, \quad (7)$$

$$Y_j^n = \sum_i \sum_m T_{ij}^{mn}, \quad (8)$$

where  $X_i^m$  is the total production of factor  $m$  in zone  $i$ ,  $Y_j^n$  is the total consumption by factor  $n$  in zone  $j$ , and  $T_{ij}^{mn}$  is the trade of factor  $m$  from zone  $i$  to be consumed by factor  $n$  in zone  $j$ .

As explained at the beginning of Section 3.1, the values of the factors in the trade matrix are normally expressed in monetary units, but they can also be expressed in physical units that can be observed and measured (e.g. tonnage or passengers transported from zone  $i$  to  $j$ ).

Given the trade matrix, it is possible to estimate the spatial relationships between zones by establishing a matrix of coefficients  $b_{ij}^m$  that represents the input of factor  $m$  produced in zone  $i$  for consumption in zone  $j$  by all factors  $n$ . These coefficients are called trade coefficients, i.e.

$$b_{ij}^m = \frac{T_{ij}^{mn}}{\sum_n Y_j^n a^{mn}}, \quad (9)$$

where  $b_{ij}^m$  is the spatial relationship between zone of production  $i$  and zone of consumption  $j$  for factor  $m$ ,  $T_{ij}^{mn}$  is the trade of factor  $m$  from  $i$  to be used or consumed at zone  $j$ ,  $Y_j^n$  is the total consumption by factor  $n$  in zone  $j$ , and  $a^{mn}$  is the functional relationship between factor  $m$  and  $n$  (see eq. (5)).

In standard multiregional input-output models these trade coefficients are fixed through time, which means that the pattern of trade between zones is also fixed. Clearly this is highly unrealistic, and trade coefficients should be adjusted through time to represent the different prices of factors among zones as well as the conditions of trade between zones, e.g. transport costs and other barriers to trade such as import duties. The trade coefficients for a factor  $m$  can be estimated as a function of the trade costs or generalized transaction costs between zones using a multinomial logit model. The probability of trade for factor  $m$  from each zone  $i$  to a particular zone  $j$  is proportional to the trade cost between zones  $i$  and  $j$ , i.e.

$$b_{ij}^m \propto \exp(-\beta^m c_{ij}^m), \quad (10)$$

where  $b_{ij}^m$  is the trade coefficient (probability) for factor  $m$  produced in zone  $i$  and consumed in zone  $j$ ,  $\beta^m$  is the elasticity of trade for factor  $m$  with respect to trade costs, and  $c_{ij}^m$  is the trade (or generalized) cost for factor  $m$  between zones  $i$  and  $j$ .

The trade cost should include the following components:

$$c_{ij}^m = p_i^m + t_{ij}^m + w_{ij}^m, \quad (11)$$

where  $c_{ij}^m$  is the trade (or generalized cost) for factor  $m$  between the zone of production  $i$  and the zone of consumption  $j$ ,  $p_i^m$  is the price of factor  $m$  at production zone  $i$ ,  $t_{ij}^m$  is the transport cost for factor  $m$  between the zone of production  $i$  and the zone of consumption  $j$  (which should include all transaction costs including fares, cost of time, duties, etc.), and  $w_{ij}^m$  is all other non-monetary costs that influence the pattern of trade of factor  $m$  between zones  $i$  and  $j$ .

Weights  $w_{ij}$  are more difficult to observe than price  $p_i$  and transport costs  $t_{ij}$ . Normally, the weights are estimated by calibrating them to reproduce the observed pattern of trade for a factor  $m$  between zones  $i$  and  $j$ . The resulting weights are expressed in equivalent monetary units, and are necessary because the aggregation of products and services into modeled factors masks differences of specialized products and services that are required for production or consumption in certain zones. The weight in zone  $i$  can also reflect the economies of scale at the production zone. Also, sometimes, when the systems modeled represent inter-regional or international trade, it is essential to take historical factors into account, in order to explain the pattern of trade between countries or regions.

The fundamental trade equation used for all transactions in space can now be formulated as follows:

$$T_{ij}^{mn} = \frac{Y_j^m \exp(-\beta^m c_{ij}^m)}{\sum_i \exp(-\beta^m c_{ij}^m)}, \quad (12)$$

where  $T_{ij}^m$  is the trade of factor  $m$  between zone of production  $i$  and zone of consumption  $j$ ,  $Y_j^m$  is the consumption of factor  $m$  in zone  $j$ ,  $c_{ij}^m$  is the trade cost for factor  $m$  between zones  $i$  and  $j$  (see eq. (11)), and  $\beta^m$  is a modifying parameter.

Note that eq. (12) is what Wilson (1970) described as a singly constrained destination model. It ensures that the consumption of  $m$  demanded in zone  $j$  is satisfied by bringing it from all possible zones of production  $i$  as a function of the relative trade costs between zones and the economies of scale in the production zones. It is also a multinomial logit model.

It is important to observe that for factors that are not transportable, such as land, the cost of transport is infinite when  $i \neq j$  and zero when  $i = j$ . In other words, the demand for the factor must be satisfied in the same zone where it is demanded.

### 3.3. Estimation of spatial prices

The output price of a factor  $n$  in a zone  $p_j^n$  can be estimated by the summation of the input prices of factors  $m$  that are required for the production of one unit of

factor  $n$ , i.e.

$$p_j^n = \sum_m a^{mn} p_j^m, \quad (13)$$

where  $p_j^n$  is the output (production) price of factor  $n$  in zone  $j$ ,  $a^{mn}$  is the technical coefficient that represents the necessary inputs of factor  $m$  to produce one unit of factor  $n$  (see eq. (6)), and  $p_j^m$  is the input (consumption) price of factor  $m$  in zone  $j$ .

The consumption price  $p_j^m$  can be calculated using the following expression (Williams, 1977):

$$p_j^m = \frac{-1}{\beta^m \log \sum_i \exp(-\beta^m c_{ij}^m)}, \quad (14)$$

where  $p_j^m$  is the consumption price of factor  $m$  in zone  $j$ ,  $\beta^m$  is the elasticity with respect to trade cost, and  $c_{ij}^m$  is the trade cost of factor  $m$  from the zone of production  $i$  to the zone of consumption  $j$ , as described in eq. (11).

It can be noticed from eq. (14) that the consumption price of the factor  $m$  in zone  $j$  is not only dependent on the trade cost but also on the parameter  $\beta$ . If there is more than one production zone from which factor  $m$  is brought to the consumption zone  $j$ , and depending on the parameter  $\beta$ , the average final consumption price could be lower than the minimum trade cost from eq. (11). At face value this counter-intuitive result would be unbelievable, but it can be interpreted as the result of competition between different producers. The model determines that when there is more than one supplier of factor  $m$ , and depending on the parameter  $\beta$ , the average consumption price is less than the minimum individual trade cost. Parameter  $\beta$  can be interpreted as the degree of market imperfection in a spatial system. If  $\beta$  is large, the market imperfection is small, as traders will choose the factor from the minimum trade cost production zone. If  $\beta$  is small, the market imperfection is large, as traders will choose from many competing production zones, because they will obtain reductions in prices from competing suppliers.

Another aspect to take into account in the determination of spatial prices is the existence of constraints that limit the production of a factor in a zone. Constraints can be real physical ones such as the amount of land available in a zone, or one imposed by regulation. Typical constraints imposed by authorities are the maximum amount of building floor space allowable or the prohibition of a type of production in a zone (e.g. zoning regulations). The consequence of the application of constraints to production is the generation of a “pure rent” in Marshallian terms, when demand for production is above the level of the constraint:

$$p_j^n = p_j^{n*} + 1/\beta^n \log(X_j^n/K_j^n) \quad \text{for } X_j^n > K_j^n, \quad (15)$$

where  $p_j^n$  is the new adjusted price of  $n$  at  $j$  when there is a constraint for production in zone  $j$ ,  $p_j^{n*}$  is the initial price of  $n$  at  $j$ ,  $X_j^n$  is the total production of  $n$  allocated at  $j$ ,  $K_j^n$  is the constraint of production of  $n$  at  $j$  (exogenous), and  $\beta^n$  is a parameter to be estimated.

### 3.4. Changing functional relationships (variable demand coefficients)

The functional relationships between factors can be considered fixed, as in traditional input-output models. A more realistic approach is to allow the functional relationship to vary depending on the prices of inputs. A number of functional forms can be utilized to estimate the variable demand coefficients. An example functional form adopted in models is described below. It refers to demand for inputs  $m$  such as food, durables, and housing space by a household of type  $n$ , and assumes a Cobb-Douglas utility function:

$$u_j^n = \prod_n (a_j^{mn} - a_{\min}^{mn})^{\alpha^m}, \quad (16)$$

where  $u_j^n$  is the utility of consumption by household type  $n$  in zone  $j$ ,  $a_j^{mn}$  is the consumption of input  $m$  by household  $n$  in zone  $j$  (see eq. (17)),  $a_{\min}^{mn}$  is the minimum consumption of input  $m$  by household  $n$  (given exogenously), and  $\alpha^m$  is a modifying parameter

The variable demand coefficient allows for the income of household  $n$  (output price of  $n$  in zone  $j$ ) to vary according to the price of inputs  $n$ . The functional form described below assumes that households of type  $n$  have an expectation of a standard of living that can be summarized by a total value of utility  $U^n$ . So, if input prices of factors  $m$  increases in a zone  $j$  to  $p_j^m$ , households  $n$  will demand a higher salary,  $p_j^n$ , to maintain the standard of living (or maintain the same overall utility), but varying the bundle of inputs consumed, i.e.

$$a_j^{mn} = a_{\min}^{mn} + \frac{U^n \alpha^{mn} / p_j^m}{\left( \prod_n \alpha^{mn} / p_j^m \right)^{\alpha^m}}, \quad (17)$$

where  $a_j^{mn}$  is the demand for input  $m$  by household  $n$  in zone  $j$ ,  $a_{\min}^{mn}$  is the minimum consumption of input  $m$  by household  $n$ ,  $U^n$  is the expected utility by household  $n$  (given),  $\alpha^{mn}$  is the parameter related to input  $m$  for household  $n$  ( $\sum_m \alpha^{mn} = 1$ ), and  $p_j^m$  is the price of input  $m$  in consumption zone  $j$  (see eq. (13)).

Using the consumption demand from eq. (17), it is possible to estimate the output price (or income demanded) by household  $n$ . It is equivalent to eq. (12), but this time using elastic coefficients that vary as a function of the price of inputs, which are zonally based, i.e.

$$p_j^n = \sum_m a_j^{mn} p_j^m, \quad (18)$$

where  $p_j^n$  is the output price of  $n$  in zone  $j$ ,  $a_j^{mn}$  is the demand coefficient for input  $m$  by  $n$  in zone  $j$ , and  $p_j^m$  is the price of input  $m$  in zone  $j$ .

### 3.5. Modeling the transport systems

Consider a transport system that moves freight and passenger flow from zones of origin to zones of destinations. A typical freight and passenger flow is described by the superscript  $s$ . Normally  $s$  indicates a type of commodity such as bulk, liquid, refrigerated, general cargo, containerized cargo, etc., or a purpose of a trip, such as journeys to work, school, shopping, recreation, business, etc. Detailed transport models include not only the cargo type or purpose of the trip but also the socio-economic group of the traveler or the commodity type.

It is possible to estimate physical flows of freight (tonnage) or passengers to be assigned to transport networks from the trade of factor  $m$  from production zone  $i$  to consumption zone  $j$  (see eq. (12)).

First, transform the trade expressed normally in annual monetary values into physical daily volumes of freight and passenger flows by applying a value-to-volume ratio for each factor, and a scalar to convert into daily flows, i.e.

$$F_{ij}^s = \sum_m T_{ij}^m v^{ms} q^m, \quad (19)$$

where  $F_{ij}^s$  is the daily flow of freight or passengers  $s$  between origin zone  $i$  and destination zone  $j$ ,  $T_{ij}^m$  is the annual trade of factor  $m$  from production zone  $i$  to consumption zone  $j$ ,  $v^{ms}$  is the ratio of value of factor  $m$  to volume of flow type  $s$ , and  $q^m$  is the conversion from year to day for trade  $m$ .

Secondly, estimate the mode of travel  $k$  at time  $t$  by route  $r$ , using a multinomial logit model, i.e.

$$F_{ij}^{ktr} \propto \sum_s F_{ij}^s \exp(-\lambda^{skt} g_{ij}^{sktr}), \quad (20)$$

where  $F_{ij}^{ktr}$  is the flow of vehicles by mode  $k$ , time of day  $t$ , using route  $r$ ,  $F_{ij}^s$  is the flow of passengers or freight type  $s$  between zones  $i$  and  $j$ ,  $\lambda^{skt}$  is the parameter for flow type  $s$ , mode  $k$ , and time  $t$ , and  $g_{ij}^{sktr}$  is the generalized cost of travel by mode  $k$  at time  $t$  and route  $r$  between zones  $i$  and  $j$  for flow type  $s$ .

The generalized cost of travel for each flow type  $s$  includes the following:

- monetary cost – includes the fare or tariff of travel between  $i$  and  $j$  such as the public transport fare, petrol and parking costs for cars, the tariff for transporting goods, tolls incurred, etc.;

- time cost – includes the time taken daily transformed into monetary units by using a value of time parameter (usually estimated empirically by revealed-preferences surveys);
- other costs – comfort and reliability that are described as mode-specific constants and expressed as extra-monetary costs for less comfortable or unreliable modes of transport (usually estimated empirically by revealed-preference surveys).

There are a number of algorithms for assigning the flow of transport to the networks. One of the most successful that is consistent with random utility theory is stochastic user equilibrium, as described by Sheffi (1985).

Once the flows are assigned to links of the network and suitably transformed into units that are the measure of link capacity, it is possible to estimate the time delays in each link (or node) of the network, i.e.

$$d_l = d_{l_0} + f(V_l/K_l), \quad (21)$$

where  $d_l$  is the time taken to traverse the link or node  $l$ ,  $d_{l_0}$  is the free-flow time (the idealized time that would be needed to traverse the link  $l$  when there are no delays due to congestion),  $V_l$  is the volume of vehicles in the link  $l$ , and  $K_l$  is the capacity of the link or node  $l$  that is measured in equivalent vehicles.

Many functions are available to estimate the delays, and are given as options in the different transport software packages.

The time of travel for each link or node of the network is aggregated into alternative routes  $r$ , at different times of the day  $t$  (e.g. peak or non-peak hours) and by different modes of transport (e.g. car, bus, truck, rail, or a combination). The aggregation uses the log sum equation that can be ‘nested’ into discrete choices at each level of a decision tree (Williams, 1977). The resulting average cost is fed back into eq. (20) to estimate flows, and into eq. (11) to estimate the trade cost that determines the spatial interaction.

### *3.6. Integrated spatial system model*

A system model is illustrated in Figure 3 that comprises a set of elements and a set of relationships. The elements to be modeled are the locations of production ( $X_i^m$ ) and consumption ( $Y_j^m$ ), and the relationships between them expressed in the spatial interaction trade matrix ( $T_{ij}^{mn}$ ).

Given a stimulus to the system of either an increase or decrease of production or consumption, the system will respond, depending on the parameters that govern the relationships between elements. These parameters represent elasticities with respect to prices, as explained above. The model thus represents a spatial market at equilibrium, when prices of the elements or factors stabilize in all zones.

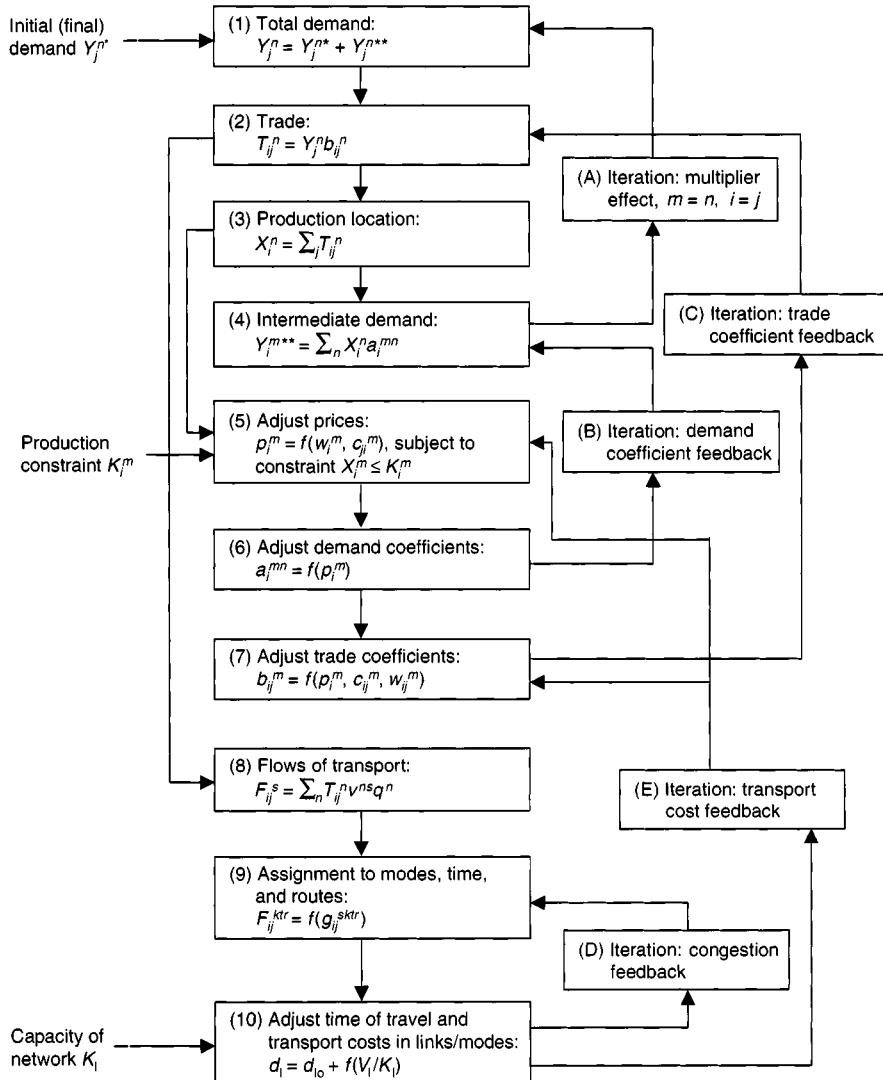


Figure 3. An integrated spatial system.

- *Step 1.* To initiate the model it is necessary to exogenously input a change in the final demand. That is to say, exports from the area modeled, investment, or government expenditure. The model estimates the total demand for factor  $j$  in each zone by adding the final or initial demand to the intermediate demands calculated in step 4. At initiation, the intermediate demand is equal to zero.
- *Step 2.* The trade of factors from production to consumption is estimated using a trade coefficient determined in step 7.
- *Step 3.* The production location is estimated by summing up the demands from all consumption zones.
- *Step 4.* The production of factors in a given zone requires intermediate inputs that are calculated using the demand coefficients from step 6. The new demand for intermediate factors is fed back to step 1. Several type-A iterations are required until no changes on intermediate demands are produced.
- *Step 5.* Prices are adjusted to take into account the trade cost from different zones where factors are transported from, and also the competition from different producers. In this step, it is also essential to take into account constraints to production. The most obvious one is the amount of land available, which may generate an extra price or “pure rent” for bringing demand in line with the restricted supply.
- *Step 6.* The changes in prices might produce changes in the demand coefficients.
- *Step 7.* The changes in prices might produce a change in the trade coefficients, making certain zones more attractive for purchasing the required factor. The trade coefficients will also depend on the transport costs coming from the transport assignment.
- *Step 8.* Transport flows are derived from the trade matrices obtained in step 2. The annual trade exchanges are transformed into volumes of daily freight and passengers.
- *Step 9.* The transport flows are assigned to modes of travel (car, truck, etc.) at a time period (e.g. peak/off-peak) and route alternatives. The model is a nested multinomial logit derived from random utility theory.
- *Step 10.* The transport flows assigned to the network are transformed into equivalent units by which the capacity of the network is measured. Given the volumes assigned and the capacity restriction, it is possible to adjust the time of travel in each route, time of day and work.

There are a number of feedback loops in the integrated model:

- *Feedback A.* This represents the multiplier effect of a change of final demand, and is analogous to the standard solution of an input–output model.

- *Feedback B.* Given a change of relative prices of inputs, the demand coefficients are adjusted to take into account these changes. This is particularly important for factors that are restricted in their supply at a given zone (e.g. land).
- *Feedback C.* Given the changes in prices in different zones either as a product of constraints in production or as transport costs increase, the trade coefficients are adjusted. Trade therefore will change.
- *Feedback D.* As transport flows are assigned to links and nodes of the network, congestion can creep in, changing the time of travel between zones, by different modes, time of day, and routes. This feedback is the main equilibrating mechanism of a transport model.
- *Feedback E.* The time changes affect the trade costs, which in turn affect the pattern of trade. This feedback is the essential link between transport and land use.

### 3.7. Policy modeling

The model described in Section 3.6 can be estimated sequentially using maximum-likelihood methods (Williams, 1979). Once it is properly estimated and validated against independent data, it is possible to use it for policy testing.

There are three types of policy instruments than can be tested in the model:

- *Investment policies.* Essentially, investment policies affecting the capacity of the transport network (e.g. a new road, port, or railway) and/or land development (e.g. hospitals and schools). The changes in investment are input to steps 5 and 10 by changing capacity constraints.
- *Pricing policies.* Decision-makers can impose taxes or subsidies on prices. The policies can be input in step 5, affecting production or consumption prices, and in step 10, affecting transport costs.
- *Regulatory policies.* Decision-makers can impose restrictions to the use of land or transport. For example, a zoning regulation that forbids industrial production will have an effect on prices and thus trading patterns. Equally, authorities impose restrictions in the use of a network such as bus only lanes, or pedestrian zones, which will affect the transport costs and thus trading. The policies can be input in steps 5 and 10 by modifying the capacity constraints.

The combination of any set of instruments described above may represent a plan or program, and their impacts can be assessed by the model.

Assessment methods, such as cost-benefit analysis, can be used for evaluating the policies. This normally consists of forecasting the state of the system modeled with and without the policies. The resulting quantities and prices for each option

produced by the model can be compared, and thus the costs and benefits of each alternative can be estimated.

#### **4. Applications**

Models are simplified for practical use, depending on the purpose of the application and the availability of resources. In some cases the emphasis is put on inter-urban movement of freight and passengers as well as on the regional location of economic activities. In other cases the emphasis is put on intra-urban movement and the location of households and jobs. It may depend also on the nature of the application. For example, in the study of the viability and the impact of a large infrastructure investment such as the tunnel under the English Channel or the development of a high-speed train network in Spain, the emphasis was put on inter-regional or international trade. On the other hand, where the interest was in developing an expansion plan of a city such as Cambridge in the UK, or the renewal of an industrial city such as Bilbao in Spain, the emphasis was shifted to model in detail the urban land use market and the movement of passengers.

Resources vary from study to study. Lack of disaggregate data may impair the extent of the details modeled. Limited time availability may lead to the simplification of the models used. In fact none of the models in practice utilizes the full model as described above, but modeling software can cope with a complete model if so desired.

Modeling software has been available for a number of years that is flexible enough to model a variety of information ranging in scale from cities to regions or countries, and also allowing flexibility for the user to disaggregate some factors and aggregate others, depending on the purpose of the application. One such modeling software is MEPLAN (Marcial Echenique, 1992); another is TRANUS (de la Barra, 1989).

One of the earliest applications at the regional level of part of the model described above was in the São Paulo State study in 1975. The model described in Williams and Echenique (1978) had an emphasis on inter-urban freight and passenger transport, and utilized fixed-demand coefficients, but variable trade coefficients, in an input-output framework. The model provided the basis for elaborating the first state transport plan. The framework has been further developed in a number of studies at the inter-regional and international scales. Among these, it is worth noting the modeling of the regional economic impacts of the Channel Tunnel (Rohr and Williams, 1994). A model for the central region of Chile (Echenique et al., 1994) has been used for planning the infrastructure of that region.

On an urban scale, the earlier application of the framework, but with emphasis on urban land uses and passenger transport, was in Bilbao (Geraldes et al., 1978).

The model has been in continuous use for evaluating highway and underground investments, as well as several major studies of urban renewal. The forecast of patronage of a new metro line in Bilbao has proven to be accurate. Anas (1984) developed a similar model in the USA, firmly anchored in discrete choice theory. A development of the same model framework has been used in the UK for London since 1988, for a variety of land use and transport planning purposes. Similarly to Bilbao, the use of the London model has proven to be fairly accurate in estimating the changes in office rents as a result of a new metro line and also forecasted well the impact of the introduction of the recent congestion charge scheme in central London. Similar models have been used in Naples (Hunt, 1994) and Sacramento (Abraham and Hunt, 2000).

## 5. Conclusions

The models described in this chapter provide a practical tool for transport and land use planning at different scales. The models have been utilized for over 25 years to forecast the impacts of transport and land investment, regulation, and pricing policies for urban areas and regions.

The models have evolved from *ad hoc* foundations to become more solidly grounded in economic theory – which provides a better explanation of the behavior of firms and households within a spatial economy than earlier approaches. The outcomes of the models at equilibrium forecast the location of firms and households and trade between them in space. The models also estimate the prices paid for all factors of the economy. Given the forecast of the quantities and prices, it is possible to evaluate in a sound manner (e.g. cost–benefit) the impacts of policies affecting the spatial economy.

The key elements that relate the equilibrium model to practical land use and transport planning is the application of constraints to production and trade. These constraints reflect the physical supply of land and infrastructure that, together with regulatory or pricing decisions, are the instruments that planners can use for guiding the spatial economy.

## References

- Abraham, J. and J.D. Hunt (2000) "Policy analogies using the Sacramento MEPLAN land use transportation infrastructure model," in: *80th Annual Meeting of the Transportation Research Board*, Paper. Washington, DC.
- Anas, A. (1984) "Discrete choice theory and the general equilibrium of employment, housing and travel networks in a Lowry-type model of the urban economy," *Environment and Planning A*, 16:1489–1502.
- Batty, M. (1976) *Urban modelling: algorithms, calibrations, predictions*. Cambridge: Cambridge University Press.

- Bröcker, J. (1995) "Chamberlinian spatial computable general equilibrium modelling: a theoretical framework," *Economic Systems Research*, 7:137-149.
- de la Barra, T. (1989) *Integrated land use and transport modelling*. Cambridge: Cambridge University Press.
- Domencich, T. and D. McFadden (1975) *Urban travel demand: a behavioural analysis*. Amsterdam: North Holland.
- Echenique, M. (1994) "Urban and regional studies at the Martin Centre: its origins, its present, its future," *Environment and Planning B*, 21:517-534.
- Echenique, M., Y. Jin, J. Burgos and A. Gil (1994) "An integrated land-use/transport strategy for the development of the central region of Chile," *Traffic Engineering and Control*, Sept.:491-497.
- Geraldes, P., M. Echenique and I.N. Williams (1978) "A spatial economic model for Bilbao," in: *Proceedings of the PTRC Summer Annual Meeting*. London: PTRC.
- Hunt, J.D. (1994) "Calibrating the Naples land use and transport model," *Environment and Planning B*, 21:569-590.
- Lee, D.B. (1973) "Requiem for large scale models," *Journal of the American Institute of Planners*, 39:163-178.
- Leontief, W.W. (1951) *The structure of the American economy 1919-1939*, 2nd edn. Oxford: Oxford University Press.
- Lowry, I.S. (1964) *A model of metropolises*. Santa Monica: Rand.
- Marcial Echenique (1992) *Technical Introduction to MEPLAN*. Cambridge: Marcial Echenique.
- Rohr, C. and I.N. Williams (1994) "Modelling the regional economic impacts of the Channel Tunnel," *Environment and Planning B*, 21:555-568.
- Sheffii, Y. (1985) *Urban transportation networks*. Englewood Cliffs: Prentice Hall.
- Webster, F.V., P.H. Bly and N.J. Paulley, eds (1988) *Urban land use and transport interaction: policies and models*. Aldershot: Gower.
- Wegener, M. (1994) "Operational urban models," *Journal of the American Planning Association*, 60:17-29.
- Williams, H.C.W.L. (1977) "On the formulation of travel demand models and user-benefit measures," *Environment and Planning A*, 9:285-344.
- Williams, I.N. (1979) "An approach to solving spatial-allocation models with constraints," *Environment and Planning A*, 11:3-22.
- Williams, I.N. and M.H. Echenique (1978) "A regional model for commodity and passenger flows," in: *Proceedings of the PTRC Summer Annual Meeting*. London: PTRC.
- Wilson, A.G. (1970) *Entropy in urban and regional modelling*. London: Pion.

*Chapter 13*

## INTRODUCTION TO URBAN SIMULATION: DESIGN AND DEVELOPMENT OF OPERATIONAL MODELS

PAUL WADDELL and GUDMUNDUR F. ULFARSSON

*University of Washington, Seattle, WA*

### 1. The context and objectives for urban simulation

Urban systems are becoming ever larger and increasingly complex as urban economies, social and political structures and norms, and transportation and other infrastructure systems and technologies evolve. Scarce resources make efficiency critically important, and in a democratic context that involves many stakeholders with conflicting values and priorities, it is neither feasible nor appropriate to deal with major land use and transportation policies and investments as isolated choices to be decided by planners or bureaucrats within the bounds of a single organization.

Mathematical and theoretical models have long been used to attempt to reduce complexity and encode a clear and concise understanding of some aspects of urban structure and transportation, as exemplified by the classic work on the monocentric model of the city (Alonso, 1964; Mills, 1967; Muth, 1969). While the value of theoretical models is facilitating a broad understanding of some underlying principles of urban development and transportation, much of this work remains too simplified in its assumptions and too abstract to be of direct value to agencies needing to inform decisions about specific policies and investments in particular urban settings.

To begin to address more operational needs in planning and policy decisions, computerized models representing urban travel and land use began to be developed and used from at least the 1960s in the USA, with the advent of the urban transportation planning system for travel demand forecasting (Weiner, 1997), and the subsequent work on spatial interaction models for predicting locations of households and jobs across urban landscapes (Putman, 1983), which emerged out of earlier work on the Lowry gravity model (Goldner, 1971). A separate branch of applied urban modeling developed along the lines of the input–output model of the macro-economy, created to describe the structure of economic flows between economic sectors (Leontief, 1966), adding a spatial component and

transportation costs to represent economic and transport flows between zones in a region (de la Barra, 1989; Marcial Echenique, 1995).

The objectives for much of the work on land use and transportation modeling in the USA from the 1960s through the 1980s were focused on the planning problem of determining transportation capacity needs – mostly focusing on roadway capacity – to accommodate expected demand generated by predicted land use patterns represented by the spatial distribution of households and jobs within a metropolitan area at some future planning horizon. During the 1970s and 1980s, increasing pressure from environmental groups, proponents of transit, and others, led to a substantial shift in US policy objectives, reflected in the passage of the Clean Air Act Amendments (CAA) of 1991 and the Intermodal Surface Transportation Efficiency Act (ISTEA) of 1990.

By 1990, a significant degree of attention had emerged on the effects of transportation improvements on land use changes, the potential for long-term induced demand from highway expansion that might significantly undermine the expanded capacity through additional travel, and increasing environmental consequences in the form of emissions and loss of open space due to stimulation of low-density development at and beyond the urban fringe. The passage of the CAAA and ISTEA legislation set the stage for lawsuits by the Sierra Club and the Environmental Defense Fund and other environmental groups in the San Francisco Bay area, Chicago, Salt Lake City, and other metropolitan areas in the USA, on the grounds that the computerized transportation and land use modeling and planning processes did not adequately account for these complex feedbacks between transportation improvements, land use, and air quality (Garrett and Wachs, 1996).

Simultaneously, the policy environment began to shift toward a more multi-modal approach to transportation, including non-motorized and transit modes, other demand-side policies began to emerge as alternative ways to match capacity to needs, including a range of transportation system management techniques (ramp metering, traffic light signalization, and so forth), travel demand management (ride-sharing, staggered work hours, parking pricing policies, congestion pricing, etc.), and land use policies (jobs-housing balance, urban growth boundaries, transferable development rights, concurrency requirements or adequate public facilities ordinances). The range of policies and strategies under potential consideration by metropolitan areas to address transportation needs has essentially exploded over the past two decades, from a fairly narrow focus on highway capacity expansion to a multi-modal transportation capacity and demand management and land use policies. The objectives for operational urban land use and transportation models have consequently grown.

Besides the growing need to test the effects and effectiveness of an ever-more diverse range of land use and transportation policies, and their interactions, pressures on operational modeling have grown from a very different perspective.

From the earliest efforts to develop operational urban models, critics have raised serious concerns about the viability of such models. Lee (1973) cogently argued that efforts to develop operational urban models had failed, and would likely continue to fail for a variety of reasons ranging from insufficient theory to computational and data demands. Much of the operational work in urban transportation and land use modeling has been criticized as being too much akin to a “black box,” meaning that its theory and implementation were not clear enough to an observer attempting to understand and evaluate it. While some of the criticism was aimed at problems that have since clearly been addressed, such as computational requirements, other concerns, such as insufficient behavioral theory, still require substantial attention and are not widely addressed even in many current operational simulation models. It is valuable to keep these critiques in mind when examining current and emerging modeling approaches.

Combined with this kind of skepticism on a technical level, growing pressures have emerged on the planning and policy-making arenas to become more open and participatory – in short, more democratic. The tradition that has emerged within planning agencies of having technical staff run models to support planning processes, without clear and open access to the models, their assumptions, their theoretical foundation, and their practical application, has become very inconsistent with the current context demanding more democratic analysis and decision processes.

In summary, the context and objectives for urban modeling have grown far more complex over the past two decades, and combine to shape the needs for urban model development in ways that are sensitive to a range of land use and transportation policies and their interactions, that build on clear and defensible foundations in behavioral theory, and that facilitate participation in the testing of alternative policy strategies and their evaluation. These are formidable challenges to address in a satisfactory way.

## **2. The design and implementation of an operational urban simulation system**

The remainder of this chapter explores recent advances and experience in the design and development of urban models that attempt to address the requirements and contextual factors described above. The general questions of model design and application discussed below are grounded in a case study of the development of the UrbanSim system in the Puget Sound region, and the rationale for each design choice is presented. UrbanSim has been developed since the late 1990s to address many of the concerns identified above, and represents an ongoing interdisciplinary research development effort to provide operational tools to support the assessment of land use, transportation and environmental

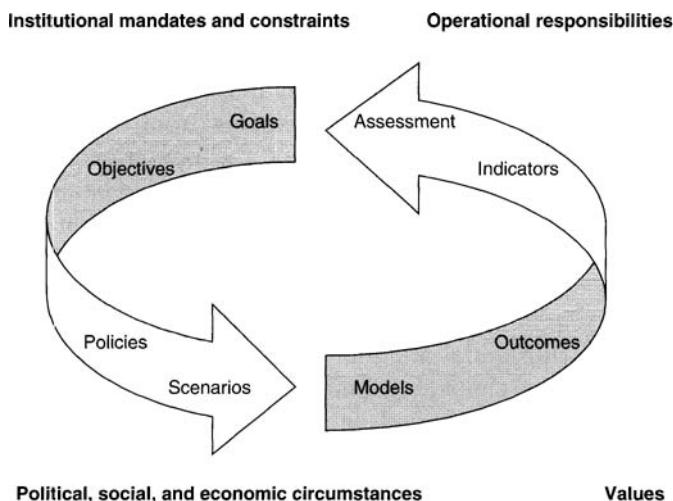


Figure 1. Models in the policy process.

policies, and plans within metropolitan areas (Waddell, 2000, 2002; Noth et al., 2001; Waddell et al., 2003).

It is proposed that models be considered within a broader context in which they will be used to guide or inform policy choices, and that this be considered a participatory and iterative process. Figure 1 depicts the proposed policy development process as one that begins with a visioning, or goal-setting phase, and proceeds through development of objectives, identifying policies, formulating policy packages or scenarios, using models to examine the effects of these policy scenarios on important outcomes, and developing indicators and evaluating the effectiveness of the policy scenarios in achieving the original policy goals and objectives. The process is likely to be iterative for several reasons, chiefly that different stakeholders will disagree about the relative weight to place on each goal, and there may be many possible policy scenarios that could be evaluated. Ultimately, the process should lead to a convergence of agreement on a set of goals and on the preferred policy strategy for achieving them. Our hope is that a well-designed policy process that integrates use of models in a participatory decision process will increase the likelihood of a cooperative resolution, compared with the frequently observed political gridlock now observed in many metropolitan regions.

The model proposed development process is summarized in the following steps:

- (1) assess the institutional, political, and technical context;
- (2) assess the stakeholders, value conflicts, and public policy objectives;

- (3) develop measurable benchmarks for the objectives;
- (4) inventory the policies to be tested;
- (5) map the policy inputs to outcomes;
- (6) assess the model requirements;
- (7) prepare the input data;
- (8) develop the model specification;
- (9) estimate the model parameters;
- (10) calibrate the model system;
- (11) validate the model system;
- (12) operational use.

Since the focus of this chapter is on the design process for developing an operational urban simulation model system, steps 1–8 are covered in substantial detail and but only a brief summary of steps 9–12 is provided.

## *2.1. Assess the institutional, political, and technical context*

Who will be using the model system, and who will be affected by its use? What are the institutional mandates and limitations of the organizations involved? What technical requirements or limitations impose bounding conditions on the problem at hand? These and related questions logically precede any model development exercise, and set its broad scope and direction.

The Puget Sound region in the state of Washington, USA, will be used as a case study for clarifying these questions and in the discussion that follows. The federally-designated metropolitan planning organization (MPO) and state-designated regional transportation planning organization (RTPO) for the region, which contains the major cities of Seattle, Tacoma, and Bellevue, is the Puget Sound Regional Council (PSRC). The PSRC coordinates the development of a regional transportation plan, which is updated every 3 years, with a major update involving significant model applications every 6 years.

In 2000, the PSRC commissioned a study to evaluate their current land use and transportation models, and to develop a long-term development strategy for new model development. The results of this effort are documented in a series of reports (Waddell et al., 2001a,b,c). The major planning responsibilities of the PSRC are summarized in Table 1, and although the focus of the organization is heavily oriented toward regional transportation planning, it also serves as a regional coordinator of land use plans due to the adoption of a state Growth Management Act (GMA) in 1990. The PSRC has no direct taxation or operational power other than these planning functions, and local governments still retain full control of land use policies and most transportation investments. The agency is therefore similar to MPOs in the USA in having relatively little political authority

**Table 1**  
**Major planning responsibilities of the PSRC**

Responsibility	Comments
Transportation	<p>Produce a regional transportation plan (RTP) that will establish the planning direction for regionally significant transportation projects</p> <p>Establish regional transportation policy and set minimum standards for the state government to integrate into its transportation planning</p> <p>Carry out MPO functions (federally mandated):</p> <ul style="list-style-type: none"> <li>• preparation of an RTP (20 year plan for integrated regional transportation system with both short and long-term actions)</li> <li>• an annual work program</li> <li>• collaborative planning program with Washington State Department of Transportation (WSDOT), transit operators, and air quality agencies</li> <li>• 3 year transportation improvement program</li> </ul> <p>Carry out RTPO functions (state mandated):</p> <ul style="list-style-type: none"> <li>• preparation of an RTP</li> <li>• 6 year capital plan</li> <li>• certify that transportation elements of local comprehensive plans are consistent with the regional transportation plan</li> <li>• certify that transportation elements of county, city and town comprehensive plans are consistent with state comprehensive planning law</li> <li>• ensure that the region's transportation projects are consistent with the RTP</li> <li>• manage right-of-way preservation proposals for high-capacity transportation development, in conformance with the RTP and other regional strategies</li> <li>• work with WSDOT to plan corridor transportation strategies</li> </ul> <p>Determine categories of priorities for the region among recommended regionally significant transportation projects</p> <p>Review and comment in the federal and state environmental impact assessment process (NEPA/SEPA) on proposed actions with potential significant impact on the implementation of the RTP</p>
Growth management	<p>Maintain Vision 2020 (Puget Sound Regional Council, 1995) and updates as the regional growth management strategy</p> <p>Develop multi-county planning policies</p> <p>Coordinate local and regional growth management planning efforts</p>
Countywide comprehensive plans	Review all countywide plans for consistency with the adopted regional growth and transportation strategy

Table 1  
Contd

Responsibility	Comments
Regional database development	<p>Support development of the RTP and regional growth management</p> <p>Forecast and monitor economic, demographic, and travel conditions</p> <p>Develop a database jointly with relevant state agencies</p> <p>Respond to data prepared by the Washington Office of Financial Management</p>
Technical assistance	<p>Provide technical assistance to local and state governments</p> <p>Provide general planning assistance to small cities and towns</p>
Discussion forum	Provide a forum for discussion among local and state officials and other interested parties

Source: PSRC Interlocal Agreements 1981 and 1998 as summarized in Waddell et al. (2001c).

or leverage other than through its role as a regional planning and coordinating agency.

This institutional context influences the planning and design of an operational urban simulation model. First, it implies a regional scope to the modeling in order to support the primary responsibility of the agency to develop the regional transportation plan. Second, it implies that there should be a significant degree of involvement and coordination with local cities and counties within the region if the PSRC is to be able to leverage a well-coordinated and effective set of land use and transportation policies and investments. Third, due to the dispersion of political authority at a local level, especially for land use policies, it suggests that the model system should be designed in a way that is useful to local governments in developing and evaluating land use and transportation policies, so that these policies may be more effectively coordinated across the region.

A significant aspect of urban planning and policy in the US context, particularly in the Pacific Northwest, is the degree of involvement by community and advocacy groups in the planning process. In Portland, Oregon, a well-documented process was spearheaded by grass-roots environmental activists who opposed a proposed circumferential highway around Portland that they argued would lead to further low-density development and more auto use. The Land Use, Transportation and Air Quality Connection (LUTRAQ, 1993) process involved environmental groups and other community advocates in the metropolitan planning process, and led to the development of alternative policy scenarios that were more transit

oriented and involved the use of land use policies to concentrate development around transit stations. Ultimately, in this case, the scenario proposed by LUTRAQ proved to be persuasive, and the highway project was abandoned.

In the Puget Sound, there are many non-profit agencies representing advocacy interests, such as protecting the environment (Sierra Club and Northwest Environment Watch), promoting non-motorized transportation alternatives (Transportation Choices Coalition), promoting growth management (1000 Friends of Washington), and arguing against the regional transit agency (Sound Transit) light rail plans (Sane Transit) and for road expansion (Citizens for Mobility), and organizations promoting limitations in governmental taxing authority and spending (Permanent Offense). In this populist context, in which the process often receives as much attention as the outcome, the development of a model system for regional planning of transportation investments and land use policies must be attentive to the active role of public participation, and the many diverse stakeholders and values that must be recognized as an integral part of the process of planning and decision-making in the region.

## *2.2. Assess the stakeholders, value conflicts, and public policy objectives*

As should be clear from the foregoing discussion, there are many stakeholders involved in regional planning and decision-making in the Puget Sound region as well as in metropolitan areas elsewhere in the USA and abroad, and that they often hold conflicting values over community priorities. Some of these conflicts arise from NIMBYism (not in my backyard), a pattern of resistance to location decisions in which parochial self-interests weigh against broader regional objectives. Many other conflicts are over more fundamental differences in values, such as how important it is to move cars faster on the roadways as compared with preserving forest and agricultural lands, or increasing the affordability of the housing in the region, or promoting economic growth. In the Puget Sound, efforts to bring these diverse interests and perspectives together to develop a consensus around a long-term vision for the region were coordinated by the PSRC, culminating in a strategy called Vision 2020 (Puget Sound Regional Council, 1995). This process will be updated in 2004, with a new target of 2030. In other regions, similar visioning efforts have been carried out by non-profit organizations or coalitions of public and private organizations, such as the Envision Utah process in the Greater Wasatch Front region of Utah. Vision 2020 presented a consensus on four major themes for policy development (the following excerpts are from Puget Sound Regional Council, 1995):

- *Improve efficiency through effective transportation system management.* The strategy places the highest priority on maintenance and preservation of all

elements of the transportation system: roads, transit, ferries, freight and goods, and non-motorized.

- *Use transportation demand management measures to reduce travel demand, provide new sources of revenue, and help meet environmental objectives.* In the short term, the region will pursue incentives to encourage transit use, ride-sharing, bicycling, walking, and telecommuting. In the long term, the region will study and consider transportation pricing strategies to generate revenue for system improvements, provide incentives for travel outside of peak periods, and discourage the growth of driving alone.
- *Focus transportation investments to support transit and pedestrian-oriented land use patterns.* Serving compact communities with high-quality transit service and locating bus stops near residences are examples of effective ways to reduce the need for motor vehicle use.
- *Add transportation capacity where appropriate to provide alternatives to automobile travel, enhance safety and access, and improve freight and goods mobility.* The strategy stresses the importance of system planning for non-motorized and transit facilities and services, so that these enhancements can be programmed similar to street and highway improvements rather than occurring in a piecemeal fashion. Improvements to the road network to provide a more comprehensive and connected roadway system are also called for.

The value of a coordinated and participatory visioning process such as Vision 2020 or Envision Utah is that it elicits from a diverse set of stakeholders a broad sense of shared values that can help set the stage for developing politically viable strategies for action, and avoiding political impasse or legal confrontation. Increasingly, the environmental movement in the USA has found a viable strategy in legally confronting MPOs and state departments of transportation over planned highway projects, arguing under the Clean Air Act Amendments that the secondary effects of the highway projects on land use and air quality were not adequately assessed, and the courts have frequently agreed. Several lawsuits around the country have documented that this strategy can be successful in blocking major highway projects. In order to avoid such legal impasse, regional visioning and planning efforts must reach out to environmental groups and other stakeholders representing a range of diverse values and perspectives. The implications for model development are significant. Models will be heavily questioned by skeptical stakeholders, and should be made as transparent in their design as possible. Ideally, the full range of stakeholders should be involved in the process of designing and implementing models in order to ensure that the design avoids any significant biases that favor one stakeholder perspective over another. This would decrease the likelihood of a legal challenge being made in the first place, and decrease the likelihood that such a legal challenge would be successful. The recent emergence of a value-sensitive design methodology in information

technology can be productively applied to the development and design of complex urban simulation models (Friedman and Kahn, 1994; Friedman et al., 2002).

### *2.3. Develop measurable benchmarks for the objectives*

Once policy objectives are identified, even while there remain significant differences in perspective over their relative priority, it is possible and quite useful to begin to develop measurable benchmarks for evaluating progress toward these objectives. If there are thresholds that are relevant due to federal, state, or local requirements, such as for air quality, these provide clear and measurable targets for achieving policy objectives. In other cases, such as transportation efficiency, it may be much more difficult to obtain consensus around a set of benchmarks that should be attained, or even what measures should be used to represent progress. For example, there may be significant differences of opinion over what levels of service on the roadways are acceptable, or whether multi-modal levels of service should be used instead. There is also little agreement about whether mobility (efficiency in moving from one location to another) or accessibility (ease in reaching desired activities) should be the guiding principle for transportation policy. These two approaches yield very different strategies for intervention, with the former focusing on expansion of roadway capacity and the latter on mixed modes and coordinated land use policies.

Unfortunately, this phase of planning is too often either ignored or left ambiguous, providing insufficient support for later stages in the process. A preferred approach would be to establish clear benchmarks where possible, and at the least develop clear measures that indicate the direction of progress on each objective. These measures should then form the basis for indicators that are used in evaluating alternative policy strategies after developing and implementing the model system and using it to compare a baseline and alternative policy scenarios. This topic will be returned to later in the description of the process.

### *2.4. Inventory the policies to be tested*

The inventory of policies in Box 1 is an incomplete but representative list of the types of policies under consideration in many metropolitan areas that affect or are affected by land use and transportation choices. Many other policies could be potentially included for analysis, and each one would require an assessment of its impact on the model design. Some policies may impose significant costs on the model development effort in order to effectively address them, but fall fairly low on a prioritization of policies to test. In such cases, it may be appropriate to drop the policy from further consideration, recognizing this as a design choice.

**Box 1**  
Policies to be potentially evaluated

**Transportation capacity**

- Expansion of roadways
- Expansion of fixed-guideway transit systems
- Expansion of bus transit systems
- Expansion of high-occupancy vehicle lanes
- Expansion of bicycle and pedestrian facilities and amenities

**Transportation system management**

- Highway ramp metering
- Incident response systems
- Traffic signalization
- Traffic-calming measures

**Transportation demand management**

- Incentives for car-pooling/van-pooling
- Incentives for staggered work hours
- Parking pricing
- Congestion pricing

**Land use/growth management policies**

- Urban growth boundaries
- Concurrency requirements or adequate public facilities ordinances
- Comprehensive land use plans
- Zoning
- Promoting urban designs such as neo-traditional neighborhood design
- Transit-oriented development
- Transferable development rights

**Incentives for infill and redevelopment**

- Development impact fees

**Economic development policies**

- Property tax abatements, incentives for business or real estate development

**Environmental policies**

- Protection of environmentally sensitive or hazardous areas
- Air quality conformity measures

## 2.5. Map the policy inputs to outcomes

For each of the policies to be considered, there is some *a priori* expectation, or range of expectations, about how the policy would affect outcomes that are intended, and others that may be indirect effects. The conceptual mapping of policy inputs to outcomes should be informed broadly by theory within the social and natural sciences. The relevant theoretical foundations for this activity cannot adequately be described within the scope of this chapter, and the reader is referred to other sources to elaborate on the treatment here (e.g. DiPasquale and Wheaton, 1990; Waddell and Moore, 2001).

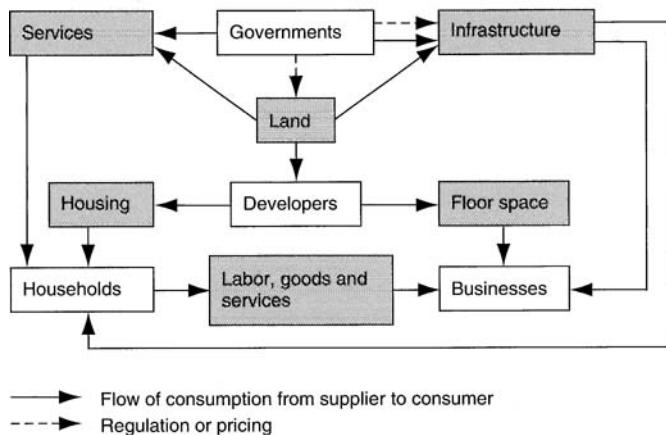


Figure 2. Linked urban markets.

Figure 2 portrays at a general level the broad scope of interactions among households, firms, developers, and governments within markets for real estate, labor, and goods and services. Developers use land to construct housing and non-residential floor space that are demanded by households and businesses, who are also interacting in the labor market and the markets for goods and services. Governments provide infrastructure and services, regulate, and, in some cases, alter prices for the use of land and infrastructure. This general framework provides a point of departure for considering the effects of alternative governmental policies and investments.

The key agents that generate or respond to the policies outlined are households, individuals, employers, developers, and governments. Households make a cluster of interdependent long-term lifestyle choices, including when to move, neighborhoods to locate within, the type of housing to rent or purchase, and the number of vehicles to own (Salomon et al., 2002). Individuals within households choose their labor force and educational status, their job mobility and job search, their daily activity schedule, and their transportation mode and route. Employers choose to start and close establishments, and choose site locations, size of employment, and the types and quantities of real estate to rent or purchase. Developers choose to undertake real-estate development projects, and the scale and locations of those projects. Governments set policies and make investments that affect the choices of other agents, and also make development choices regarding public facilities, including type, location, and scale of development.

The agents, choices, and interactions suggested as appropriate to connect a broad range of policy inputs to outcomes are summarized in Figure 3. It

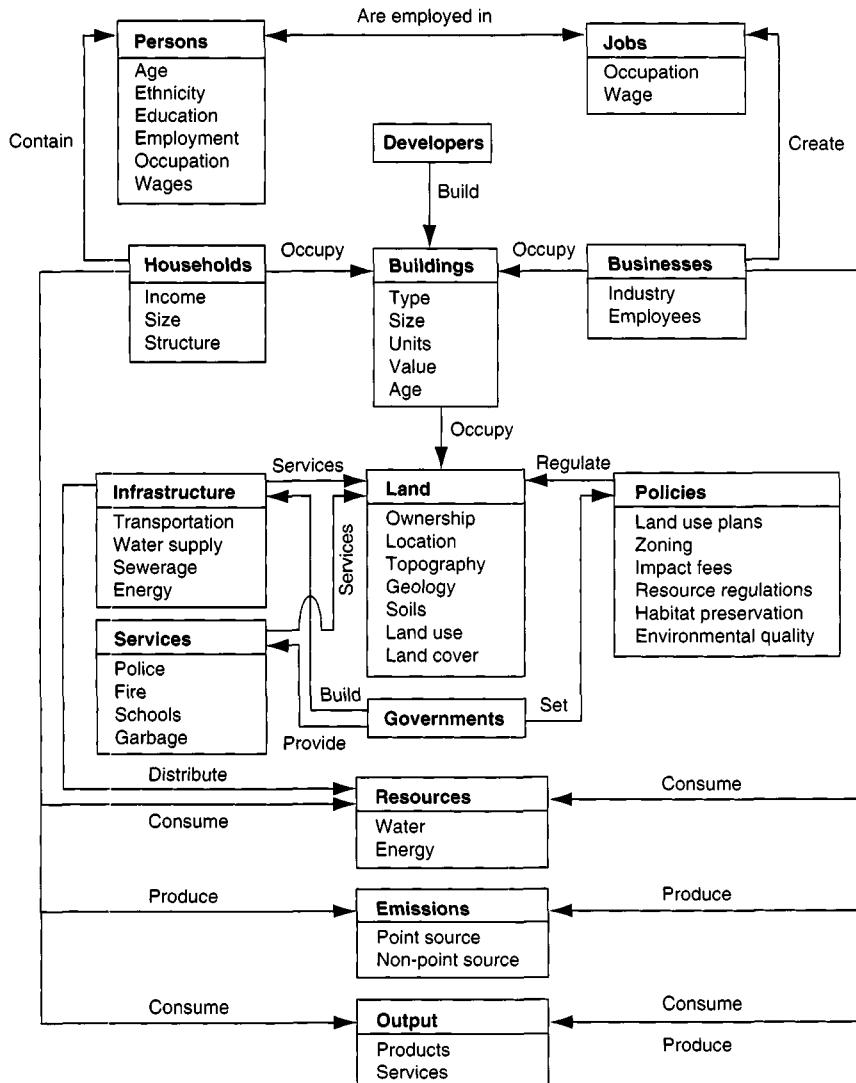


Figure 3. Agents, choices, and interactions to represent in a complete urban model.

is suggested that governmental actions such as regulations and infrastructure investments be treated as exogenous, if the objective of the modeling is to evaluate alternative public policies. The other choices and processes depicted lead to

logical structures for model components, and the representation of agents and objects on which they act.

For each of the policies identified, it should be possible to trace the expected causal paths from policy input to outcomes, and this conceptual mapping should provide a foundation for developing a model design that will be responsive to the policy in question. Although there is insufficient space available to do this systematically for those policies listed in Box 1, one example is presented, and the reader is left to consider how this might be applied to other policies.

Consider the example of the imposition of an urban growth boundary (UGB). The intent of such a policy is to encourage compact development within the boundary, and limit urban development outside the boundary, in order to protect farm and resource lands from urbanization and to promote more efficient use of existing infrastructure. Critics argue that it also leads to rapid housing price inflation. How, and to what extent, does the UGB produce these outcomes? First, it should be noted that the UGB is actually not a direct regulatory policy that is implemented independently of other policies. It is actually a higher-order policy that must be implemented through changes in the comprehensive land use plans of cities and counties affected by the boundary, in order to make those plans consistent with the intent of the UGB. In other words, these local jurisdictions may be required by state law first, to delineate a UGB, and second, to change their land use plans and zoning to "downzone" the areas outside the UGB to an agricultural intensity, and to "upzone" the areas within the UGB in order to increase the intensity of land use and to accommodate anticipated development over a planning horizon. In Washington, the UGB is intended to include sufficient developable area to accommodate 20 years of development, and is to be revisited and extended when it no longer contains sufficient area to accommodate this level of anticipated development.

In practical terms, the intent of the UGB policy is operationalized through land use plans and zoning and coordinated with other infrastructure choices and land use policies, so the model must be made sensitive to these policies if it is to be sensitive to UGB policies. Second, the effects of the UGB policy on prices are uncertain, since there are competing forces at work. On the one hand, the UGB policy and the resulting "downzoning" of areas outside it clearly limit the supply of land for future development. On the other hand, the UGB is to be delineated so that it contains 20 years of development potential, and there are counter-balancing "upzoning" policies within the UGB. So the effect of the policy on prices is theoretically ambiguous, and the model would need to be sensitive to the effects of a relative scarcity of land for development, at different levels of intensity of zoning. Furthermore, there may be indirect effects on household location choices, with the UGB creating a scarce amenity of access to open space for those that would live near the interior border, or in the existing housing that was developed outside the UGB before the policy went into effect.

## 2.6. Assess the model requirements

A review of background documents provided the foundation for much of the institutional context for model development in the Puget Sound (Waddell et al., 2001c). This was augmented by an extensive review of the literature on operational land use and transportation models (Waddell et al., 2001a). Based on these materials and on extensive interviews with PSRC staff, and surveys of staff in local governments, the Washington State Department of Transportation and several advocacy groups, the following requirements emerged for new land use and transportation model development at the PSRC (Waddell et al., 2001b).

The model system must:

- Be sensitive to the effects of transportation pricing policies on both travel behavior and land use.
- Be able to address the impacts of proposed land use and transportation policies on housing affordability.
- Be able to support the role of the regional council in monitoring development and compliance with the objectives of Vision 2020 and the Growth Management Act.
- Be sensitive to policies that are designed to promote densification, infill, and redevelopment, and to the effects of land use policies such as comprehensive land use plans, zoning, and the UGB on real estate development and the location of households and firms.
- Be able to assess policy effects over periods ranging from less than 5 years to 30 years.
- Be designed to support a participatory policy process that includes activities such as Vision 2020, where scenarios are generated and publicly discussed. Public access to the model assumptions, theory, structure, and results is required, and the models must be explainable to a non-technical audience.
- Support the creation of performance indicators and evaluation measures suitable for use by the regional council in evaluating alternative policy scenarios using, at a minimum, least-cost planning and cost–benefit analysis techniques.
- Be based on an activity-based framework in order to adequately represent the complexity and constraints of travel behavior, and the influence of land use and transportation policies on travel behavior.
- Allow comparison of different transit modes, for example rail versus bus. The new model system must be capable of adequately representing non-motorized travel behavior. The model must allow comparison of different auto modes of travel, for example high-occupancy vehicles with two or more persons versus single-occupancy vehicles.
- Recognize the impact of land use patterns on demand for transportation.

- Allow analysis of demand induced by transportation system improvements.
- Be able to assess the impacts of environmental regulations that affect the development of environmentally sensitive lands, such as salmon habitat, wetlands, floodplains, seismic areas, and steep slopes.
- Be able to assess the impacts of commute trip reduction and transportation-demand management policies, such as different work arrangements (flexible-versus fixed-work schedule, telecommuting or not, compressed work week or regular work week).
- Recognize the effects of multi-modal transportation systems and policy changes on real estate development and the location patterns of households and firms.
- Be sensitive to the effects of urban design elements such as mixed land use, density, street pattern, transit service, and pedestrian amenities on household and firm location and travel behavior.
- Be able to analyze the residential movement and location choices made by households, and the influence on these choices of relevant housing and location characteristics. The model system must be able to model the choice of household members to participate in the labor market, and to choose a work location. The model system must be able to model the vehicle ownership choices of households.
- Be able to analyze the interactions between household choices related to residential mobility and location, labor market participation and workplace, vehicle ownership, and daily activity and travel scheduling.
- Be able to represent demographic processes such as the change in household size and structure, and the aging of the population.
- Incorporate a component to model the process of real estate development, including infill and redevelopment, and the effects of various policies on this process. The new models must distinguish between important types of real estate that are relevant to the goals and objectives of Vision 2020, including adequate representation of different non-residential, residential, and mixed-use types.
- Be able to support the analysis of transit-oriented development, including real estate development, and household and business location, to assist in station area planning.
- Incorporate a macroeconomic component to model economic growth in the region and its relationship to internal and external economic drivers; and analyze the factors and policies influencing the location choices and real estate demands of different firms in different industries.
- Address freight and commodity transport within and through the region.
- Address modal choices and trade-offs of moving goods by truck, rail, barge, or air.

- Be able to produce multi-modal travel assignments for roadway, transit, and, possibly, non-motorized systems.
- Contain information pertinent to commute management systems policies such as workplace incentives, telecommuting, and a greater breakdown of carpool sizes. The model should provide information showing the effects of transportation system management initiatives.
- Provide output relevant to policies instituted as part of intelligent transportation systems such as incident management systems or public information distribution.
- Be developed in a way that supports open and unrestricted access to the software by regional council staff, consultants, and constituents, in order to maintain and modify the models to meet emerging needs over time. The model system should support distributed access and use by regional council member governments, and should use consistent data for both regional council and member agency applications.
- Be manageable from the perspective of its data requirements.
- Have reasonable performance, in terms of computational efficiency, so that an entire run of the full land use and transportation model system can be accomplished within one working day.
- Provide tools to facilitate visualization of the model results and comparison of scenarios in ways that are useful to non-technical audiences. The model must provide forecasts of system characteristics tracked by state benchmarks, for example, the benchmarks identified by the state of Washington's Blue Ribbon Commission on Transportation. The model system must produce output that allows analysis of the best mix of transportation investments.
- Allow multi-modal cost–benefit analysis, to enable model users to make more informed transportation investment decisions.
- Address uncertainty in the models and produce ranges of values for outputs rather than specific results to avoid suggesting artificial accuracy.

## 2.7. *Make preliminary model design choices*

Having reviewed the context, policy applications, and requirements for model development, the stage is set for examining the design choices in more detail. The remainder of this chapter explores the design choices and process for developing an operational urban simulation model that satisfies the requirements discussed in the preceding sections. As design options are discussed, the choices actually made in the design of UrbanSim to address such requirements will be explained, and the discussion will alternate between general topics of model design and specifics of these design choices within UrbanSim.

There are several major design choices that must be considered for an urban simulation system, and the combination of these choices narrows the choice of modeling approach. The design choices considered here are the level of behavioral aggregation, the level of determinism, their temporal representation, and the resolution of agents, space, and time.

### *Behavioral resolution*

The system can work on an aggregate scale of average behaviors or on the disaggregate level of behavior of individual agents. The simulation system can be deterministic or stochastic. Deterministic systems are based on predetermined rates of change and static functions, and the same inputs will always produce the same results. Deterministic models are typically used along with an aggregate scale of behavior, since the average behaviors can often be approximated with a fixed rate of change. But, agent-based simulations can also easily be deterministic, for example queuing models, or choice models that force the agents to choose the most likely result.

### *Resolution of agents, space, and time*

The size of the units of analysis, or resolution, of the system is another major design decision. Simulation systems range from macroscopic to microscopic in resolution. Macroscopic systems have the largest units of analysis, typically aggregate values for geographic zones, household distributions, or groups of vehicles. These models are in wide use, mostly because they have relatively low computational needs. They run relatively fast and use relatively little memory. Macroscopic models also require much fewer data than finer resolutions, and the data are more readily available through census data or other such large databases. Macroscopic models are typically static and deterministic.

Microscopic models have a small unit of analysis, for example a single individual, household, vehicle, trip, or activity. Microscopic models are being actively developed because their requirements for considerable computational power are increasingly being satisfied by personal computers, making these models more feasible, and they support clearer behavioral specifications than macroscopic models. Microscopic models are typically stochastic, disaggregate models that require enormous amounts of data. The data needs of microscopic models are still a limitation since detailed data on individuals are expensive to collect. Methods to synthesize households from census data exist (Beckman et al., 1995), and such methods facilitate the creation of synthetic households for use in micro-simulation.

Mesoscopic models are a mixture of macroscopic and microscopic models. They may use small decision-makers but large time steps, or small time steps for

large units of analysis. They therefore allow the use of aggregate data for certain aspects of the model but make use of greater detail where it is available.

### *Level of determinism*

Stochastic models are based on probability distributions. They are most typically used with agent-based simulations in the form of probabilistic choice models. They allow the agent to randomly choose an alternative, which means an agent can end up with an unlikely choice. These models will generally not give the same result if run twice, but the results can be made repeatable by fixing the random seed that controls the random distribution. Such a feature does not make the model deterministic, since the outcome is always probabilistic if the initial condition is changed slightly.

### *Temporal representation*

Models can also be cross-sectional or dynamic in their representation of time. Cross-sectional (sometimes referred to as static or equilibrium) models are not time-dependent, and model conditions are fixed at a hypothetical condition generally identified as a long-term equilibrium. Equilibrium models assume that the system begins in equilibrium and adjusts completely to some exogenous shock, that is, it reaches a new equilibrium. The assumption of equilibrium usually allows the explicit derivation of the solution describing the system, although the underlying functions may be time-dependent. Dynamic models make no assumption about equilibrium, but concentrate on adjustment processes over real calendar time. The solution describing such a model is therefore often impossible to derive analytically and the system must typically be simulated to find the result.

### *System interaction*

There are complex interactions between components within the urban system that must be represented in any complete urban simulation system, such as the endogenous relationships between land use, transportation, and the environment. Urban simulation systems must therefore either explicitly model the interactions between land use, transportation, and the environment, or interface with separate transportation or environmental models.

This interaction, or interface, between systems is made especially complex because of the different time-scales of urban development, transportation, and environmental changes. In particular, there is a large contrast between urban development and transportation. Simulations of urban development work on time-scales from a year down to a month, while transportation simulation systems are on a scale of days down to seconds. Environmental models can be on both

scales, for example the effect on wildlife habitat works on the urban development time-scale, i.e. years or months, while models of pollution will be on the transportation model scale.

## *2.8. Select the modeling approach*

The modeling approach is a major design decision, though it will be heavily constrained by the preceding design choices. Several different urban modeling approaches have been developed and applied to either planning or research objectives over the past several decades, and after a hiatus of almost two decades in the USA there is now an active and rapidly growing array of research activities developing and deploying new modeling approaches. Three of these methods were used in the earliest operational urban models, dating from the 1960s and 1970s: spatial interaction, spatial input–output, and linear programming. Microsimulation was developed in the 1960s, but not applied to urban modeling until the 1980s. Since the 1980s, the development of discrete choice modeling and the emergence of cellular automata (CA) and multi-agent simulation (MAS) techniques have created a proliferation of modeling approaches. We discuss each of these approaches below, and the supporting role of geographic information systems (GIS) and the integration of several of these approaches in the design of UrbanSim. The review specifically does not cover theoretical advances such as those that have arisen in the form of the new economic geography (Fujita et al., 1999), which have not been incorporated into operational planning models that are the focus of this chapter.

### *Spatial interaction*

Models based on the spatial interaction approach include some of the earliest efforts to model systematic spatial patterns of urban land use. The approach draws on the model of gravity in physics, which indicates that gravitational pull increases in proportion to the mass of two objects in space, and decreases with the square of the distance separating them. Applied to urban settlements and travel, the gravity model implies that travel between two zones increases with the amount of activity in the origin and destination zone, and decreases with the square of the travel impedance between them. The basic model has been extended to model trip destination choices, residential location choices, and employment location choices (Putman, 1983). This type of model tends to be limited in the degree of spatial detail used, and does not represent many behavioral factors influencing location choices, nor does it represent the role of real estate markets and prices.

### *Spatial input–output*

The spatial input–output framework extended the input–output model developed to represent the structure of the US economy (Leontief, 1966) to address spatial patterns of location of economic activity within regions, and the movement of goods and people between zones. For examples of complete urban models based on this approach, see de la Barra (1989) and Marcial Echenique (1995). Zones are treated in a sense as economies that engage in production, consumption, import, and export within the zone and with all other zones in the model. These economic exchanges between zones are denominated in monetary units, and driven by demand for exports. Monetary flows are converted to flows of goods and services by type of vehicle, and of commuting and shopping trips by mode. The approach includes explicit real estate and labor markets, as well as travel demand modeling, and is structured to generate a static equilibrium solution to changes in one or more inputs. Zone sizes in operational applications tend to be large relative to zone sizes used in typical urban travel modeling.

### *Linear programming*

Linear programming models of land use are rare. TOPAZ (Dickey and Leiner, 1983), developed in Australia, and POLIS, applied in the Bay Area (Prastacos, 1985), are examples of models that use this approach. Linear programming optimizes a global objective function, such as consumer surplus or utility, across the entire model system. The approach is therefore more suited to exploration of alternative land use configurations that might optimize transportation flow, than to reflect realistic behavioral responses to changes in the transportation system or in land use policies.

### *Microsimulation*

Microsimulation as an approach essentially implies a model that is applied at the level of the individual. Developed in the late 1950s and early 1960s, the method was initially applied to study the effects of social and economic policies (Orcutt, 1957; Orcutt et al., 1961). It has more recently been applied to the formulation of urban models such as MASTER (Mackett, 1992), DORTMUND (Wegener, 1985), and UrbanSim (Waddell, 2002). Microsimulation models used for socio-economic (non-spatial) policies such as taxes are used with a sample of households, and compute the effects of the tax policy alternatives on the sample of households to study distributional, or equity, effects. Urban, spatially explicit models have used combinations of discrete-choice models and transition rates to predict changes in the state of individuals or households, such as entering or leaving the labor force, and their choices such as residence location. Once

considered too data-intensive, these methods have been growing in popularity because they allow modeling at an individual level, where behavioral theory is clearer, and due to increased interest in detailed characteristics of households for equity analysis and other reasons, can make individual-level analysis more efficient than cross-classification of households using multiple characteristics.

#### *Discrete choice*

Discrete choice modeling techniques are widely used in travel demand modeling, mostly in the analysis of mode choice. Discrete choice models have been used in some form for much of the last 100 years. However, it was Daniel McFadden's Nobel prize winning random utility theory work and his derivation of the generalized extreme value class of models (which includes multinomial and nested logit models) that gave these models a firm foundation within econometrics, and they have since become standard methods in developing models that attempt to predict individual choices among a finite set of alternatives (McFadden, 1973, 1981, 1984). Discrete choice models are generic in the sense that they do not impose overly restrictive assumptions on the choice process, and have been shown capable of addressing large and complex choice sets effectively (Ben-Akiva and Lerman, 1985). Discrete choice techniques can be readily used in conjunction with other simulation approaches, such as microsimulation.

#### *Rule based*

Several land use models have been developed in recent years using GIS and a rule-based set of procedures to allocate population, employment, and/or land use. Examples include the CUF model (Landis, 1994), Uplan (Johnston, 2003), and WhatIf? (Klosterman, 1999). Such rule-based applications may have a useful role in making models more accessible, but there is a risk that model users would interpret the models as having a more behavioral basis than their rules actually contain. There are also rule-based methods that are emerging from the field of artificial intelligence, using observed data to generate clarification trees of behavioral rules that are used in microsimulation, such as the Albatross activity-based travel model (Arentze et al., 2000).

#### *Cellular automata*

CA models have emerged within the broad field of complex systems as a means of representing the emergent properties of simple behavioral rules applied to cells within a grid (Wolfram, 1984). The approach has now been widely applied to urban land cover or land use change (White and Engelen, 1993; Benati, 1997; Couclelis, 1997; Clarke and Gaydos, 1998). To date, applications have been

principally for research purposes rather than operational planning or policy, though efforts are underway to make these models useful for planning purposes. The approach is particularly useful for representing the interactions between a location and its immediate environment, but tends to reflect a fairly abstract representation of agents, decisions, and behavior, since the models focus on simulating the change in state of individual cells. In addition, challenges remain in reconciling the emergent behavior of cells acting on localized rules with more systemic or macro-scale behavior, in validating these models using observed data, and in computational requirements. Potentially the most ambitious use of the CA approach to date is the TRANSIMS traffic microsimulation system, which has been recently tested in Portland, Oregon (Los Alamos National Laboratory, 2002).

### *Multi-agent simulation*

MAS is related to CA in that both draw on complex systems theory, but differs from CA in that its emphasis is on emergent system behavior arising from interactions between agents. Research and testing of MAS models accelerated rapidly after the SWARM software environment was developed for implementing models of this type (Swarm, 2002). The MAS approach is gaining substantial research interest across the social sciences, since it opens new avenues to analyze social behavior from an interactive perspective. In economics, the adoption of MAS has come to be known as agent-based computational economics (Tesfatsion, 2000).

### *Supporting role of GIS*

The growing use of GIS to automate land records and collect environmental data has led to substantial interest in using the technology in urban modeling. Most of the applications of GIS are of two forms: integrating the input data for use in urban models, or visualizing results of the simulation in a map-based display. While both of these roles are valuable, they do not exhaust the potential applications of GIS technology. A number of efforts have emerged to operationalize models within a GIS software environment. An example of this for travel demand modeling is the TransCAD system (Caliper Corporation, 2002).

### *Integration of modeling approaches in the design of UrbanSim*

Several of the preceding modeling approaches have been assimilated in the design and development of the UrbanSim system. It uses microsimulation to model individual choices of households and jobs. Discrete choice modeling is used to predict location choices of households and jobs, and the real estate development

choices of developers. A GIS is used to integrate input data and to display model results. Many of the computations made by UrbanSim involve spatial analysis, integrated into the model system infrastructure. By using a cell-based representation of land, and a probability of change in development type from one year to the next that is influenced by the state of neighboring cells, the real estate development model component parallels models using CA. Unlike CA models, however, UrbanSim reflects specific agents (developers) interacting with other agents (households, jobs, and governments) within a simulation environment, which reflects aspects of MAS, though in UrbanSim the granularity of interactions is presently at the model component level rather than at an individual agent level. Only two models in operation adopt a dynamic, path-dependent approach: UrbanSim, which operates at a microsimulation level, and Delta, which is aggregate in its implementation (Simmonds, 1999).

## *2.9. Prepare the input data*

Simulation systems, especially microsimulation, require enormous amounts of detailed data. A large part of any simulation project is therefore the collection of data and the preparation of that data into a form required by the simulator. The constraints of available data often influence the choice of model design, though these constraints are rapidly changing. There has been a historical tendency to assume that more aggregate data was likely to be less prone to errors than disaggregate data, but this has not been shown empirically. Moreover, when problems are detected in aggregate data, the potential for correcting data errors may be far less than with the original source data, such as parcel records.

Given the representation of the agents, choice processes, and interactions depicted in Figure 3, and the assessment of model requirements, the data needed for model development can be more clearly defined. It is clear that a reasonably complete urban simulation model will need to represent in its database land and real estate, including land use, housing, and non-residential real estate, the value of real estate, households, and their characteristics and location, jobs, and their industry and location, as well as locational references for planning areas such as local jurisdictional boundaries, travel modeling zones, and environmental features that might influence development or be influenced by it.

Parcel data are the most logical and widely available form of source data for representing information about land and real estate. It is generally available from tax assessor's offices in the USA, and is increasingly available in a GIS database with parcel boundaries, due to rapid automation of US land records. These data, even where available in electronic form, still contain errors and gaps that require attention, for example with systematic underreporting of information about properties that are tax-exempt.

Detailed data on individual households is more difficult to obtain, but census data can be used to synthesize individual households (Beckman et al., 1995). This synthetic approach involves using the iterative proportional fitting algorithm (Deming and Stephan, 1940) to estimate the joint distribution of household characteristics for census tracts or block groups, by using the correlation structure of the US Public Use Microdata Sample (PUMS) and the marginal distribution of household characteristics as given by the US census Summary File 3A tables.

Data for businesses establishments can be obtained either from governmental records collected for unemployment insurance purposes, or from commercial sources. These data, like parcel records, may contain random errors and systematic gaps, such as the underreporting of governments, schools, and self-employed proprietors. In addition, there may be reporting problems, with employment being listed at a headquarters or accounting office location for a multi-establishment firm rather than at each branch facility.

Data for planning and environmental characteristics are generally available in GIS form, and may be readily integrated into the database for model development. Environmental features such as wetlands, floodplains, steep slopes, water bodies, and sensitive habitat such as riparian buffers are useful to include in the database, since these often are features that affect land development policies that lead to constraints on urban development.

The integration of the database used in the application of UrbanSim is depicted in Figure 4, which shows the inputs and the resulting database consisting of three primary tables: grid cells, households, and jobs. The problems of integrating spatial data as listed here are sizable, and it is beyond the scope of this chapter to fully describe either the difficulties or reasonable strategies for overcoming them. GIS and database technology, as well as data mining and statistical techniques such as multiple imputation of missing data (Schafer, 1997) are making these challenges of assembling a robust, micro-level integrated urban database for urban simulation much more manageable than in the past.

## *2.10. Develop the model specification*

Given the data, the individual model components of the simulation system must be specified. This is typical, since an urban simulation system contains so many interacting agents and processes that it is impossible to specify a single joint model of them all. It would be impossible to estimate parameters for a single model encompassing all the agents and choices depicted in Figure 3, given currently available data and modeling technology. It is therefore necessary to separate the models into reasonably distinct components, especially considering the deep endogenous relationship that can affect a set of choices.

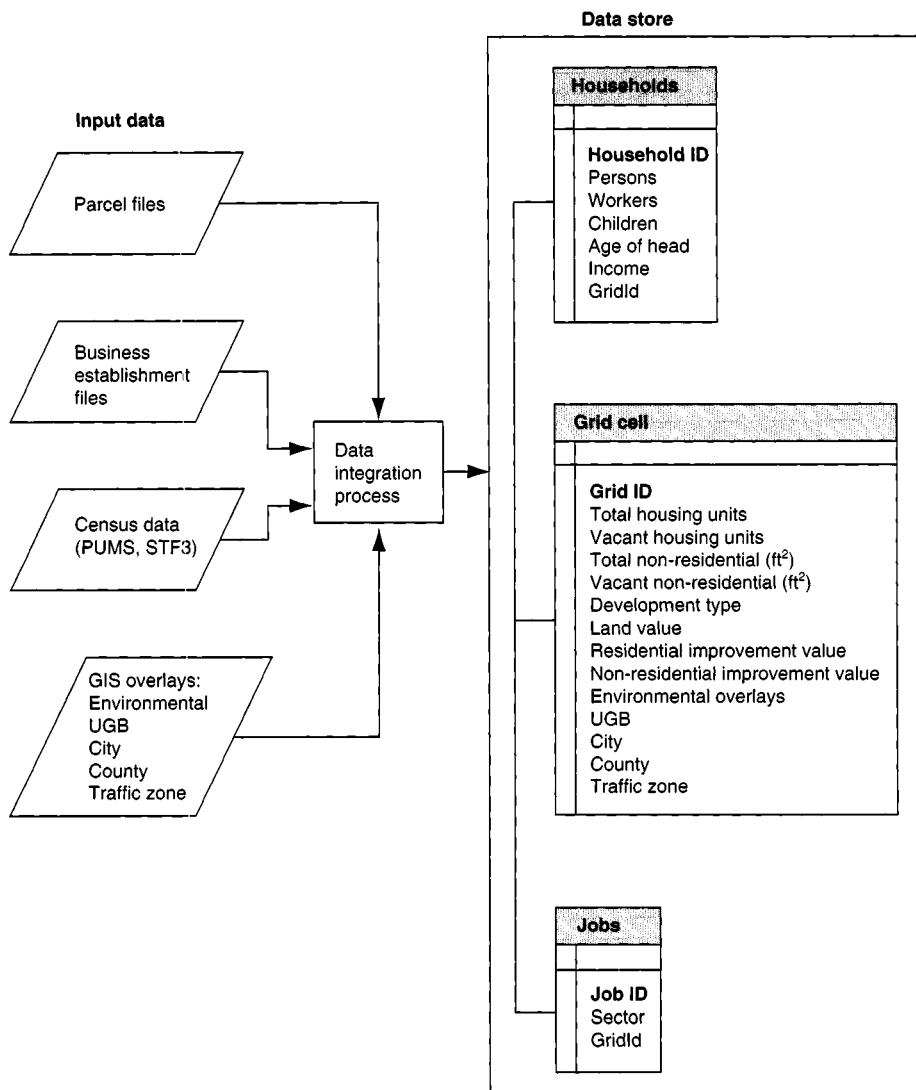


Figure 4. The UrbanSim data integration process. (Source: Waddell, 2002.)

There are many plausible ways that the models could be structured as modular and interacting components. The approach taken in the design of UrbanSim is to represent the model system as a set of interacting models for demographic transition, economic transition, household relocation, employment relocation, household location choice, employment location choice, real estate development, and land prices. Each of these is described briefly below.

- *Demographic transition* – interfaces with exogenous (external) information from macroeconomic models that predict the total population, and potentially other aggregate information about population characteristics such as household size and income distributions. The model component compares these anticipated totals for future years to the UrbanSim household database, to determine how many households of each type must be added or removed from the database in order to be consistent with the external assumptions about total growth or decline of that type of household over the period of one year.
- *Economic transition* – serves the same function as the demographic transition component, and adds or removes jobs from the UrbanSim database to achieve consistency with the external assumptions about economic growth or decline in each economic sector over the period of one year.
- *Household relocation* – predicts the probability that a household currently located in the region will move over a period of 1 year. It is sometimes referred to as “residential mobility.”
- *Employment relocation* – predicts the probability that a job in a given sector and location will be moved from that location during a year. It is sometimes referred to as “employment mobility.”
- *Household location choice* – examines all households that have been added by the demographic transition model, and those that have been selected to move within the current year by the household relocation model, and predicts their location choice from available (existing and vacant) housing units.
- *Employment location choice* – examines all jobs that have been added by the economic transition model, and those that have been predicted to move by the job relocation model, and for each job chooses a location from the available job spaces.
- *Real estate development choice* – predicts the probability that each location will experience a real estate development event over a 1 year period, given the characteristics of the location and the market conditions, and if a development event is predicted, predicts the type of development that would occur.
- *Land price* – uses location and real estate characteristics to predict land prices at each location, which then informs the location choices of households, firms, and developers in the subsequent time period.

The specification of each of these model components involves the choice of the form of the dependent variable, the functional form of the model, and the independent variables to be included in the model. A brief compilation of the dependent and independent variables, and the functional form of the key model

**Table 2**  
Specification of UrbanSim model components

Variable	Household location choice	Employment location choice	Real estate development choice	Land price
Dependent variable	150 m grid cell	150 m grid cell	Development event	Log of land price
Functional form	Multinomial logit	Multinomial logit	Multinomial logit	Multiple regression
Independent variables				
Cell				
Land use plan			✓	✓
Housing price	✓		✓	
Housing density	✓		✓	✓
Housing age	✓		✓	✓
Neighborhood				
Distance to urban edge			✓	✓
Recent development trends			✓	
Land use mix	✓		✓	✓
Land values		✓	✓	
Jobs by sector	✓	✓		
Highways/arterials		✓	✓	✓
Regional				
Job/population accessibility	✓	✓	✓	✓
Vacancy rates			✓	✓

components in UrbanSim is shown in Table 2, with the variables used in each model identified by the presence of a symbol in the respective columns.

The model specification leads from the type of process, value, or choice to be predicted. Continuous, numeric, values are typically estimated using linear regression of some type – described more fully by Greene (2003). In UrbanSim, hedonic, multivariate, regression is used to predict land prices. Many processes can be described as categorical and unordered, for example choice processes such as location choice, development choice, mode choice, route choice, and activity choice. These models are typically specified as probabilistic discrete choice models. There exists a large number of different choice models, with the logit model or a logit model variant as the most common types (Ben-Akiva and Lerman, 1985).

### *2.11. Estimate the model parameters*

After the specification of individual models and data preparation the model coefficients must be estimated. The estimation methods are most typically a least-

squares method, the method of maximum likelihood, or probabilistic simulation in the case when a non-closed form likelihood describes the model. The least-squares methods are widely known, and handle a variety of linear and non-linear forms. Many non-linear forms can be converted to linear-in-parameters form and handled with linear regressions, but otherwise non-linear least squares can be used. In addition to single-equation models, least-squares methods exist for simultaneous systems of equations, most notably seemingly unrelated regression, two-stage least squares, and three-stage least squares. For extensive details on least-squares and maximum likelihood methods see Greene (2003).

Probabilistic models are most typically estimated with the method of maximum likelihood, if the likelihood function of the model has a closed form representation. However, for non-closed form likelihoods, methods exist to estimate the coefficients using probabilistic simulation. A popular such method is the Markov chain Monte Carlo simulation (Gilks et al., 1996), which is not only robust in the presence of complex model structures but also provides information on the structure of uncertainty in the joint distribution of the parameter estimates.

## *2.12. Calibrate the model system*

Following the estimation of the model coefficients, the urban simulation system as a whole must be calibrated. The best way to do this is to have complete data for two time periods. The system is then set up to run from the first time period as its initial condition to the second time period. The system results can then be compared with the data from that time. This allows inaccuracies to be detected and will also potentially show errors in the design. The interaction of the individual models can be calibrated at this stage as desired.

## *2.13. Develop the software application*

Before a model system can be used, it must be implemented in a software application. In many modeling projects, the software is developed through a customized process that implements a model to the exact specifications of the model and the data used in developing it, but not in a way that facilitates making changes in the data or specification. These tend to be prototype software applications that rely heavily on “hard-coding” of assumptions about model specifications and data, and are therefore not very general or reusable. Good software engineering practices can significantly increase the modularity of the software, and improve its performance, ease of maintenance and evolution to address changes in data and model specifications.

Open source licensing of software is also valuable in increasing the transparency and accessibility of the source code, and has been shown in systems such as Linux to produce extremely robust code. The UrbanSim software application is based on the Java programming language, and adopts an open source licensing approach. It is freely available on the project web site at <http://www.urbansim.org>.

Modularity is increased in the UrbanSim system by forcing all data handling by models to go through a model coordinator. Data are stored in a standard SQL database, such as MySQL, and are loaded into memory to increase performance. Simulation results are stored in the database, making the simulation inputs and results accessible from other software applications such as GIS, charting, and reporting tools.

#### *2.14. Validate the model system*

To validate the model it is necessary to run it separately on data that were not used to estimate the model or calibrate the system. To do the validation, data are needed from a time period to serve as the initial condition and from a later period to serve as a comparison with model predictions. Validation is crucial to give users confidence in the system. Goodness of fit during estimation and calibration is not proof of the quality of the model, since any model can be made to closely match given conditions. It is only when a model is used for different conditions that its predictive power is given. A historical validation of the UrbanSim model applied to Eugene-Springfield, Oregon, has been previously documented (Waddell, 2002). Practical constraints on the creation of historical data for use in validation often preclude the feasibility of historical validation of this sort, but this remains one of the most informative ways to assess the model before putting it into operational use.

#### *2.15. Operational use*

The last step is the actual operation of the model. Data are prepared for as recent a time period as possible. The model users then prepare a baseline scenario that contains the assumptions against which other scenarios will be compared. Generally, for planning purposes, the baseline scenario is a “do nothing” set of assumptions that attempts to represent the policies that are currently in place, with no major policy changes. Alternative scenarios can then be constructed that contain different assumptions regarding policies and macroeconomic conditions. Any of the policies to which the model design and specification is sensitive can be included in a scenario. For example, a scenario can include a proposed highway

expansion in some future year, or a new light rail system. Also, land policies such as land use plans or urban growth boundaries can be added, removed, or changed, in addition to zoning changes.

Since no model will perfectly predict the future, model predictions are more useful as an indication of the likely direction and magnitude of effects of an alternative scenario when compared with a baseline scenario, than for use as a set of absolute predictions about the future. In some model applications, information on uncertainty of the outcomes and of the differences between scenarios can be presented, which adds valuable information that can be used to inform policy choices.

### **3. Conclusion**

This chapter has sought to explain the context, policy applications, and major design choices in the process of developing an operational urban simulation model, with specific reference to UrbanSim as a case study in model design. It has been argued that careful design at each stage of the process is needed to make the model sensitive to the policies of principal concern, to make the data and computational requirements manageable, to make the model usable by staff and other users with appropriate levels of training, and to fit into the operational practices of the relevant organizations.

To be useful (relevant) in the policy process, model design should carefully integrate the elements discussed here into a design that fits well into a specific institutional and political context, and evolve to adapt to changing conditions. This introduction to the design process sets the stage for more in-depth discussion of specification and operational issues in model use.

The UrbanSim system is being further developed to adapt to varying data availability, different factors influencing agent choices in locations ranging from newer and rapidly growing US metropolitan areas to older US regions with a declining core, as well as issues that arise in metropolitan areas in other parts of the world. Considerable effort is now being devoted to developing environmental components of the system such as land cover change, and to developing a robust user interface and tools for visualization and evaluation of policy scenarios.

### **Acknowledgments**

This research has been funded in part by the US National Science Foundation, grants CMS-9818378, EIA-0121326, and EIA-0090832. A. Borning, A.J. Brush, J. Franklin, and S. Kim are thanked for their helpful comments on earlier drafts of this chapter.

## References

- Alonso, W. (1964) *Location and land use*. Cambridge: Harvard University Press.
- Arentze, T., H. Hofman and H. van Mourik (2000) "Using decision tree induction systems for modeling space-time behavior," *Geographical Analysis*, 32:330–350.
- Beckman, R.J., K.A. Baggerly and M.D. McKay (1995) "Creating synthetic baseline populations," in: *Transportation Research Board Annual Meeting*, Paper. Washington, DC.
- Ben-Akiva, M. and S. Lerman (1985) *Discrete choice analysis: theory and application to travel demand*. Cambridge: MIT Press.
- Benati, S. (1997) "A cellular automaton for the simulation of competitive location," *Environment and Planning B*, 24:205–218.
- Caliper Corporation (2002) *TransCAD*. Newton: Caliper (<http://www.caliper.com/tcovu.htm>).
- Clarke, K. and L. Gaydos (1998) "Loose-coupling a cellular automation model and GIS: long-term urban growth prediction for San Francisco and Washington/Baltimore," *Geographical Information Science*, 12:699–714.
- Couclelis, H. (1997) "From cellular automata to urban models: new principles for model development and implementation," *Environment and Planning B*, 24:165–174.
- de la Barra, T. (1989) *Integrated land use and transport modelling*. Cambridge: Cambridge University Press.
- Deming, W. and F. Stephan (1940) "On a least squares adjustment of a sampled frequency table when the expected marginal totals are known," *Annals of Mathematical Statistics*, 11:427–444.
- Dickey, J.W. and C. Leiner (1983) "Use of TOPAZ for transportation-land use planning in a suburban county," *Transportation Research Record*, 931:20–26.
- DiPasquale, D. and W.C. Wheaton (1990) *Housing market dynamics and the future of housing prices*. Cambridge: Harvard University and MIT.
- Friedman, B. and P. Kahn (1994) "Educating computer scientists: linking the social and technical," *Communications of the ACM*, 37:64–70.
- Friedman, B., P. Kahn and A. Borning (2002) *Value sensitive design: theory and methods*. Seattle: University of Washington (<http://www.urbansim.org/papers/vsd-theory-methods-tr.pdf>).
- Fujita, M., P. Krugman and A.J. Venables (1999) *The spatial economy: cities, regions and international trade*. Cambridge: MIT Press.
- Garrett, M. and M. Wachs (1996) *Transportation planning on trial: the Clean Air Act and travel forecasting*. Thousand Oaks: Sage.
- Gilks, W.R., S. Richardson and D.J. Spiegelhalter (1996) *Markov chain Monte Carlo in practice*. London: Chapman and Hall.
- Goldner, W. (1971) "The Lowry model heritage," *Journal of the American Institute of Planners*, 37:100–110.
- Greene, W. (2003) *Econometric analysis*, 5th edn. Englewood Cliffs: Prentice Hall.
- Johnston, R.A. (2003) *Uplan: urban growth model* 2003 [cited June 1, 2003]. (<http://snepmaps.des.ucdavis.edu/uplan>).
- Klosterman, R.E. (1999) "The what if? collaborative planning support system," *Environment and Planning B*, 26:393–408.
- Landis, J.D. (1994) "The California urban futures model: a new generation of metropolitan simulation models," *Environment and Planning B*, 21:399–420.
- Lee Jr, D. (1973) "Requiem for large scale models," *Journal of the American Institute of Planners* 39:163–178.
- Leontief, W. (1966) *Input-output economics*. New York: Oxford University Press.
- Los Alamos National Laboratory (2002) *Transportation analysis simulation system (Transims) Portland study reports*. Los Alamos: LANL (<http://transims.tsasa.lanl.gov/>).
- LUTRAQ (1993) *The Lutraq alternative/analysis of alternatives*. Portland: LUTRAQ, with Cambridge Systematics, Calthorpe, and Parsons Brinkerhoff Quade and Douglas.
- McFadden, D. (1973) "Conditional logit analysis of qualitative choice behavior," in: P. Zarembka, ed., *Frontiers in econometrics*. New York: Academic Press.
- McFadden, D. (1981) "Structural discrete probability models derived from theories of choice," in: C. Manski and D. McFadden, eds, *Structural analysis of discrete data and econometric applications*. Cambridge: MIT Press.

- McFadden, D. (1984) "Econometric analysis of qualitative response models," in: Z. Griliches and M. Intriligator, eds, *Handbook of econometrics*. Amsterdam: North Holland.
- Mackett, R.L. (1992) *Micro simulation modelling of travel and locational processes: testing and further development*. London: Transport Studies Group, University College London.
- Marcial Echenique (1995) *Use of meplan to formulate and evaluate development proposals*. Cambridge: Marcial Echenique.
- Mills, E.S. (1967) "An aggregative model of resource allocation in a metropolitan area," *American Economic Review*, 57:197–210.
- Muth, R.F. (1969) *Cities and housing*. Chicago: University of Chicago Press.
- Noth, M., A. Borning and P. Waddell (2001) *An extensible, modular architecture for simulating urban development, transportation, and environmental impacts*. Seattle: University of Washington (<http://www.urbansim.org>).
- Orcutt, G. (1957) "A new type of socio-economic system," *Review of Economics and Statistics*, 38:773–797.
- Orcutt, G., M. Greenberg, J. Korbel and A. Rivlin (1961) *Microanalysis of socioeconomic systems: a simulation study*. New York: Harper and Row.
- Prastacos, P. (1985) "Urban development models for the San Francisco region: from plum to polis," *Transportation Research Record*, 1046:37–44.
- Puget Sound Regional Council (1995) *Vision 2020: 1995 update*. Seattle: PSRC (<http://www.psrc.org/projects/vision/vision2020.htm>).
- Putman, S.H. (1983) *Integrated urban models*. London: Pion.
- Salomon, I., P. Waddell and M. Wegener (2002) "Sustainable life styles? Microsimulation of household formation, housing choice and travel behaviour," in: W. Black and P. Nijkamp, eds, *Social change and sustainable transport*. Bloomington: Indiana University Press.
- Schafer, J. (1997) *Analysis of incomplete multivariate data, monographs on statistics and applied probability*. London: Chapman and Hall.
- Simmonds, D.C. (1999) "The design of the delta land-use modelling package," *Environment and Planning B*, 26:665–684.
- Swarm, G.D. (2003) *Swarm*. Santa Fe: Santa Fe Institute (<http://www.swarm.org>).
- Tesfatsion, L. (2000) *Introduction to the CE special issue on agent-based computational economics*. Ames: Department of Economics, Iowa State University.
- Waddell, P. (2000) "A behavioral simulation model for metropolitan policy analysis and planning: residential location and housing market components of urbansim," *Environment and Planning B*, 27:247–263.
- Waddell, P. (2002) "Urbansim: modeling urban development for land use, transportation and environmental planning," *Journal of the American Planning Association*, 68:297–314.
- Waddell, P. and T. Moore (2001) "Forecasting demand for urban land," in: *Land market monitoring for smart urban growth*. Cambridge: Lincoln Institute for Land Policy.
- Waddell, P., C. Bhat, E. Ruiter, S. Bekhor, M. Outwater and E. Schroer (2001a) *Review of the literature and operational models: final report to the Puget Sound Regional Council on land use and travel demand forecasting models*. Seattle: Puget Sound Regional Council ([http://www.psrc.org/datapubs/pubs/model\\_review.pdf](http://www.psrc.org/datapubs/pubs/model_review.pdf)).
- Waddell, P., M. Outwater and C. Bhat (2001b) *Recommendations for integrated land use and travel models: final report to the Puget Sound Regional Council on land use and travel demand forecasting models*. Seattle: Puget Sound Regional Council ([http://www.psrc.org/datapubs/pubs/model\\_recommendations.pdf](http://www.psrc.org/datapubs/pubs/model_recommendations.pdf)).
- Waddell, P., E. Schroer and M. Outwater (2001c) *Assessment of model requirements: final report to the Puget Sound Regional Council on land use and travel demand forecasting models*. Seattle: Puget Sound Regional Council ([http://www.psrc.org/datapubs/pubs/model\\_requirements.pdf](http://www.psrc.org/datapubs/pubs/model_requirements.pdf)).
- Waddell, P., A. Borning, M. Noth, N. Freier, M. Becke and G. Ulfarsson (2003) "Urbansim: a simulation system for land use and transportation," *Networks and Spatial Economics*, 3:43–67.
- Wegener, M. (1985) "The Dortmund housing market model: a Monte Carlo simulation of a regional housing market," in: K. Stahl, ed., *Microeconomic models of housing markets. Lecture notes in economic and mathematical systems* 239. Berlin: Springer-Verlag.
- Weiner, E. (1997) *Urban transportation planning in the united states: an historical overview*. Washington, DC: US Department of Transportation.

- White, R. and G. Engelen (1993) "Cellular automata and fractal urban form: a cellular modelling approach to the evolution of urban land use patterns," *Environment and Planning A*, 25:1175–1199.
- Wolfram, S. (1984) "Cellular automata as models of complexity," *Nature*, 311:419–424.

*Chapter 14*

## EVOLUTIONARY APPROACHES TO TRANSPORT AND SPATIAL SYSTEMS

AURA REGGIANI

*University of Bologna*

### 1. Introduction

The modern spatial economy has a global character that is leading to important socio-economic and political changes. New transport and telecommunications flows and forms play a significant role in this through their dynamic and complex interplay with the economic and political driving forces behind globalization. In analyzing such an impact, operational and measurable indices of transport and communication patterns are necessary to map out and predict emerging trends. It could be useful in this respect to consider tools and models that have also been adopted in other disciplines. Transport and spatial modeling are already closely linked. The most widely used transport models – spatial interaction/entropy models and discrete-choice models (DSMs) – have their roots in regional economics and econometrics, and still remain the fundamental models to plan and predict transport behavior. But these are mainly static models, i.e. their mapping refers to a fixed year, and the related forecast analyses are based on assumptions of constant attitudes in the user's behavior. However, the rich variety of dynamic behavior in transport and spatial settings calls for a dynamic approach to analyses and models.

Dynamic behavior has a wide range of appearance in a space-time context, ranging from slow dynamics (e.g. changes in transport infrastructures, trade, and location patterns) to fast dynamics (e.g. changes in users' behavior, technological dynamics, financial markets). As a consequence, current transport/spatial evolution often shows forms and structures exhibiting disequilibrium, uncertainty, and sudden changes – at both the global and local levels – that are difficult to predict and evaluate. In order to understand this complex environment, spatial economic models have recently centered their analysis on novel approaches, by highlighting the need for new paths in both the theoretical/methodological and empirical contexts. In particular, dynamic system modeling has, in recent decades, become an essential part of the analysis and simulation of complex behavior or phenomena. After a period in which attention was focused on linear dynamics, in

the 1970s and 1980s there was a rapid growth of interest in non-linear dynamic systems, as witnessed *inter alia* in catastrophe theory, chaos theory, and synergetics (Nijkamp and Reggiani, 1998). These advances influenced studies from the 1990s on multi-layer, interlinked, complex behavior, by calling forth a variety of approaches based on economic dynamics (endogenous growth, evolutionary theory, externality theory, game theory, search behavior, etc.), as well as on spatial economics (network theory, complexity theory, self-organization theory, bio-ecologically based theory, etc.).

## 2. Spatial choice and processes: the role of spatial interaction models

In the last century, spatial interaction models (SIMs) were the central focus of several models and theories, with the aim of describing and predicting the analysis of spatial movements, i.e. the processes or spatial flows emerging as result of given spatial configurations. SIMs are essentially models of spatial flows, i.e. flows of people, commodities, capital, or information, from some origin  $i$  to some destination  $j$ . At one time SIMs were very popular, owing to their usefulness in studying the geography of movement, and they are still considered a relevant technique for exploring the cohesion and dispersion of activities in spatial systems. SIMs have a long history, from their first formulations that assumed analogies with Newton's law of gravity. The gravity theory, a relational theory describing the degree of spatial interaction among points through analogies with the physical world, is the initial root of SIMs. Subsequently, SIMs were shown to have a basis in entropy theory and in the utility maximizing approach, and, more recently, connections with neural networks have been developed.

For transport planning purposes SIMs are often used in travel demand analysis. These models are mainly built around the four-stage demand-forecasting modeling process. SIMs are the basic models for the analysis of trip distribution, predicting how many trips made by a person of type  $k$  and originating in zone  $i$  will terminate in zone  $j$ . SIMs are also connected with the third phase of the four-stage model, the modal split analysis. Modal split concerns the prediction of the number of trips made by a person of type  $k$  from  $i$  to  $j$  by mode  $m$ , and is usually carried out by using disaggregate models, e.g. DCMs, that have emerged from microeconomic utility theory. Their aggregate form is analytically consistent with SIMs.

### 2.1. Spatial interaction models: the analytical form

The general form of a (double-constrained) SIM is

$$T_{ij} = A_i B_j O_i D_j \exp(-\beta c_{ij}), \quad i = 1, \dots, I; j = 1, \dots, J, \quad (1)$$

where  $A_i$  and  $B_j$  are balancing factors, equal to

$$A_i = \frac{1}{\sum_j B_j D_j \exp(-\beta c_{ij})}, \quad B_j = \frac{1}{\sum_i A_i O_i \exp(-\beta c_{ij})}, \quad (2)$$

derived from the respective additivity conditions:

$$\sum_j T_{ij} = O_i, \quad \sum_i T_{ij} = D_j. \quad (3)$$

In equation (1),  $T_{ij}$  represent the total number of trips (flows of cars, telephone calls, etc.) between  $i$  and  $j$ ;  $O_i$  and  $D_j$  are the stock variables (e.g. population size and workplaces) in the places of origin and destination;  $c_{ij}$  are the interaction costs; and the term  $\exp(-\beta c_{ij})$  is the deterrence function, measuring separation effects between  $i$  and  $j$ .

Equation (1) is compatible with Newton's law of gravity. However, it can also be derived as a probabilistic approach based on statistical equilibrium concepts (Wilson, 1970). Wilson demonstrated that SIMs of type 1 can be derived by a mathematical optimization problem by maximizing an entropy function, and can thus be seen as an optimum system solution. This approach offered a macro-behavioral context to SIMs, given that entropy can be interpreted in terms of a generalized cost function for spatial interaction behavior (Nijkamp, 1975). In addition, microeconomic choice theory can offer a more interesting behavioral interpretation to SIMs.

## 2.2. Spatial interaction behavior and choice behavior

SIMs are aggregated models, i.e. they are based on observed relations for groups and travelers, or on average relations at a zonal level (Ortúzar and Willumsen, 1994). A second type of model used in spatial science is the disaggregate model (DM), based on observed choices of individuals. Compared with the former models, DMs present several advantages – the orientation toward behavioral approaches, a greater flexibility in specifying choice processes, a more effective way of testing the statistical validity of empirical results from surveys or questionnaires, a better way of including qualitative information on spatial choice processes, and a more adequate representation of the dependence of the utility of an actor on the decision of all other actors (Nijkamp and Reggiani, 1992). Further positive implications are more efficiency in data usage, use of the inherent variability in the information, and less bias due to correlation among units, and probabilistic features.

The use of DMs involves more complexity in the model design, computation, and calibration procedures. They embrace models such as microsimulation

models, conventional utility-maximization models, random-utility models, activity-based choice models, and search models. The choice for a given class of models often depends on the objective of the analysis and on the database.

In the framework of choice theory, the conventional assumption is that individuals choose rationally among all the options available to them. Individuals belong to a homogeneous population, where all members perceive utility in the same way, or to a heterogeneous population, where dispersion with respect to the mean value of the utility is large. The first case identifies the deterministic utility maximization, which can also offer further analytical interpretation – in terms of minimum cost problem – to SIMs.

In the second case we deal with the behavior of a heterogeneous population where variations (intra-individual/option variations, variations in the specification of the individual position in the space/socio-economic group, variations in the option attributes, etc.) around the mean utility are large (Domencich and McFadden, 1975). Here, a probabilistic approach is used to formalize the expected choice behavior, based on the assumption of individual random-utility maximization. To represent the choice attractiveness/probability, use is made of the concept of random utility that postulates the existence of a discrete set  $J$  of available alternatives  $j$  ( $j \in J$ ); a partition of the population into homogeneous subgroups, each sharing the same choice set and characteristics, and facing the same constraints; the existence of a set  $X$  of vectors of measured attributes of the individuals and their alternatives; and the existence of an individual, random-utility function  $u_{hj}$  – associated with each alternative  $j$  by the individual  $h$  – that has to be maximized by each individual over the choice set.

This random-utility function is represented by two components – a deterministic and a random component – so that the total utility of an individual  $h$  ( $h = 1, \dots, H$ ), with respect to an alternative  $j$  ( $j = 1, \dots, J$ ), is

$$u_{hj} = v_{hj} + \xi_{hj}, \quad (4)$$

where  $u_{hj}$  is the total random utility of alternative  $j$  ( $j = 1, \dots, J$ ) for individual  $h$  ( $h = 1, \dots, H$ );  $v_{hj}$  is the deterministic or systematic part of the utility function, which is a function of the measured attributes  $x$  of alternative  $j$  and may vary from individual to individual; and  $\xi_{hj}$  is the random part of the utility function, which reflects the idiosyncrasies and differences emerging from individual taste variations, as well as any measurement or observational errors, mis-specification, etc., made by the analyst.

Under these assumptions, the probability that an individual  $h$  chooses an alternative  $j$  is

$$P_{hi} = \text{Prob}\{(v_{hj} + \xi_{hj}) \geq (v_{hj'} + \xi_{hj'})\}, \quad j = 1, \dots, J; j' = 1, \dots, J; j \neq j'. \quad (5)$$

Equation (5) identifies the random-utility maximization, which characterizes the discrete-choice family of models. Depending on the distribution of the

random residuals  $\xi_{hj}$  in equation (5), a broad family of discrete-choice models may arise (Domencich and McFadden, 1975), which can be subdivided into multinomial logit (MNL), general extreme-value models, and nested logit models; probit models; and elimination models. The application of DCMs in spatial analysis has received considerable attention. MNL and nested logit models have been successfully applied to transportation and urban economics (Ben-Akiva and Lerman, 1985). These last two categories of DCMs show compatibility with SIMs.

The MNL model is the most popular and practical form of a DCM. It emerges by assuming that the random terms  $\xi_{hj}$  are independent and identically distributed according to a Gumbel function. Furthermore, by assuming that  $v_{hj}$  describes the average behavior of the population, we obtain the final formulation of the probability of selecting alternative  $j$ :

$$P_j = \frac{\exp v_j}{\sum_k \exp v_k}, \quad j, k \in J, \quad (6)$$

which identifies the MNL model (McFadden, 1974). The probabilistic form of the SIM, derived from equation (1), corresponds to the MNL expressions of type (6) (Nijkamp and Reggiani, 1992).

Compatibility between SIMs and DCMs has been demonstrated (Anas, 1983; Batten and Boyce, 1986; Sen and Smith, 1995). SIMs are interpreted in a behavioral context with an economic meaning, by considering SIMs as aggregate models of human behavior. These findings offer a theoretical ground to SIMs, by posing methodological questions concerning whether the inherent limits of MNL models are present in the structure of SIMs.

### 3. Non-linear dynamic processes: the logistic form

Given the “universal” law of the logit/spatial interaction form in most static models adopted so far in spatial and transportation science (see the previous section), it is not surprising to find the same “constancy” in an evolutionary complex perspective. It can be shown that the MNL approach represents the steady state of network evolution. Correspondingly, the dynamic version of the logit form leads, under particular assumptions, to the well-known (Verhulst) logistic function (Nijkamp and Reggiani, 1992), which reads – in discrete terms – as follows:

$$y(t+1) = ry(t)(1-y(t)). \quad (7)$$

Equation (7) is a “degenerated dynamic MNL” form, where  $y$  denotes the probability of choosing the first alternative (i.e. a transport mode) in a binary choice situation, and the growth parameter  $r$  is directly related to the marginal

utility function. Equation (7) belongs to the family of models (May, 1976) showing cyclical behavior for the values  $3 < r \leq 3.824\dots$ , or unstable/chaotic movements for  $3.824\dots < r < 4$ . At the bifurcation value  $r = 3.824\dots$ , a period of cycle 3 appears, giving rise – according to Li and Yorke's (1975) theorem of “period three implies chaos” – to the chaotic situation, where an uncountable number of aperiodic and periodic trajectories occurs.

In a transport mode choice situation modeled by a May equation, chaotic, and therefore unpredictable, movements may result when the growth rate assumes high values, e.g. in a dense network. On the other hand, for low growth rates, the system is stable and controllable. Hence, the relevance of detecting chaos behavior lies in the identification of unpredictable events. In a multiple-choice situation, equation (7) shows the addition of the interacting terms, in the form of ecologically-based models, such as the well-known prey-predator systems (Nijkamp and Reggiani, 1992), and unstable and chaotic/unpredictable trajectories may emerge, depending on the values of the growth parameters and initial conditions.

The logistic function (7) can be embedded in the more universal expression of a niche model – a general ecologically based model expressing the phenomenon of interspecies competition and dynamic resource utilization. This 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

$$\dot{y}_i = y_i \left( K_i - \sum_{j=1}^N a_{ij} y_j \right), \quad (8)$$

where  $y_i$  is the population of species  $i$  (e.g. transport mode,  $i = 1, 2, \dots, N$ ),  $\dot{y}_i$  is the rate of change of  $y$  over time,  $K_i$  is the carrying capacity for species  $i$ , and the coefficients  $a_{ij}$  are the interaction/competition coefficients measuring the niche overlap. The dynamic processes of the substitution of transport infrastructures and energy systems over time, as well as the introduction of technological/transport innovations, can be modeled by means of equation (8). In Figure 1, niches  $y_1$  and  $y'_1$  – e.g. the users of a certain technological/transport innovation – have the same capacity  $K_1$ . However,  $y_1$  shows a more rapid take-off, as in the metropolitan areas, owing to the greater value of its growth rate. A positive evolution then occurs when a new species ( $y_2$ ) replaces, in the short or long run, the old ones ( $y_1$ ) and ( $y'_1$ ), by exploiting new network capacities (Nicolis and Prigogine, 1977). Applications based on the adoption curves of equation (8) have been carried out with reference to the evolution and diffusion of technological innovations as well as urban dynamics (Batten et al., 1987). The capacities  $K_i$  and the coefficients  $a_{ij}$  may also embed dynamic functions. Thus, the network evolution is the result of multi-layer niche dynamics, representing the interaction among niche species, as well as among the niche capacities and/or the niche growth rates.

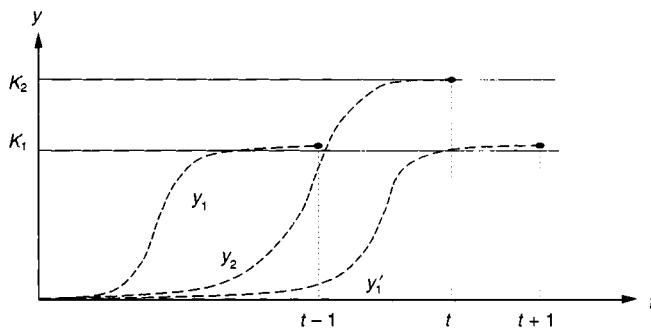


Figure 1. The logistic niches ( $y_1$  and  $y'_1$ ) occupied successively by a niche of increasing effectiveness ( $y_2$ ).

By varying the parameter, simulation experiments concerning networks of form (8) exhibit a wide spectrum of ordered, irregular, and complex behavior (Nijkamp and Reggiani, 1998). From the empirical viewpoint, such results raise the problem of prediction and control of the behavior of a complex system and, hence, the necessity of collecting quasi-dynamic or panel data on transport networks, in order to test the model coefficients, and, particularly, their speed of change. From the methodological viewpoint, this lack of predictability of future events – for certain systems – reflects a new phase in the research on spatial transport analysis. This emphasizes the need to deepen the concept of network complexity with multiple cooperative and interactive/competition effects, by exploring the links between the relatively simple behavior at the individual level and the more complex and adaptive group behavior at the meso/macro-level.

#### 4. Networks and complexity

In a transport economic context, network complexity has recently addressed the analytical/planning need to investigate the dynamic interrelationships between intermodality, interoperability, and interconnectivity, at all levels of transport/communications and organizations. Intermodality addresses the issue of the sequential use of different transport modes in the chain of transport. Interoperability refers to operational and technical uniformity that makes it possible to use and link the layers of a transport network, while interconnectivity is concerned with the horizontal coordination of, and access to, networks of different geographical coverage (Nijkamp, 1995). This led to the development of the inter-transport matrix.

Nijkamp and Reggiani (1998) argue that 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 centres) and which are connected by means of corridors and edges." Connectivity – which may be quantified by various indices – indicates the existence of multiple relationships, of alternative paths, which reinforce the interconnection of a network. Literally, the notion of a network refers to "operations via nets." The relevance of the function of the network via organized linkage patterns is embedded. The network concept, by highlighting its two inter-linked sides – its complicated morphology, and its function with stable effects – is an example of a space-time complex system that leads to synergy effects, i.e. higher economic benefits for all the actors involved. Moreover, taking into account the non-linear evolution of synergy, as well as of the multiple relationships/functions in the connected paths from a multi-layer/multi-modal perspective, we get an idea of the increased network complexity. In this context, the issue of stability and robustness, and hence of the resilience concept, become relevant for transport planning purposes.

The idea of complexity concerns the mapping of the non-intuitive behavior of a system, particularly the evolutionary patterns of connections among interacting components of a system whose long-run behavior is hard to predict. Defining complexity is fraught with difficulties. Horgan (1995) provided a list of more than 30 definitions of complexity. He particularly emphasized the relativist approach to complexity, namely from the observers' side when there is a partial understanding of the observed phenomenon (Martellato, 1998).

The objective of identifying a unified theory of complexity is still open. Systems and network theory may help in defining analytical, and hence measurable, complexity, but it remains difficult to capture inherent behavioral complexity. Mannermaa (1995) identifies the inherent complexity of models as semiotic complexity, and the inherent complexity of reality in natural processes as ontological complexity. Thus, if ontological complexity is hard to model, as also argued by Horgan, semiotic complexity is easier to analyze. One way is through the classification of Casti (1979):

- *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 type of hierarchical structure, the connectivity pattern, the variety of components, and the strength of interactions.
- *Dynamic complexity* – concerns the dynamic network behavior governed by non-linearities 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 non-linear models, e.g. chaos models, in particular, and evolutionary models in general, able to map out the dynamic (random) network patterns.

## 5. Network complexity

Modeling network complexity can, in principle, be pursued by the deductive approach when complexity emerges from the dynamic equations modeling the network concerned, and the inductive approach when complexity emerges from data analyses by means of appropriate approaches. An example of the first approach is the whole family of dynamic spatial interaction models, chaos models, etc., while, in the second typology, we find approaches detecting complexity from data analyses such as neural networks and self-organized criticality.

### 5.1. Simple models for complex networks: niche models

The role of the niche model was outlined as a unifying framework – based on the logistic form – able to map out a fully integrated network configuration with multiple nodes and links. By means of the “universal” model of form (8), evolution emerges as the result of the competition/substitution/complementarity phenomena between different subsystems/niches, leading to a self-organizing process. Keeping in mind the previous definition of “static” complexity, we can explore whether, by increasing the static complexity of a network (the number of links and nodes, an increase in the parameter values, etc.), the inherent dynamic complexity is increasing or decreasing.

In particular, it can be seen through simulation experiments (Nijkamp and Reggiani, 1998) that the simple deterministic dynamic law (8) shows the following characteristics for a complex network:

- A competing or prey-predator network, where all the elements are in a competing or prey-predator relationship, as in the transport mode network, appears to be rather robust, in the sense that the network does not collapse or explode, in the presence of high variations in the parameter values or in the carrying capacities. In addition, this kind of network shows a large spectrum of complexity patterns (regular, cyclical, and irregular behavior). In particular, the competitive network seems more robust than the prey-predator network, by allowing a large domain of the parameter values.
- A symbiotic network, where all the network elements are complementary, appears to be more fragile, in the sense that the network easily collapses or explodes, in the presence of high variations in the parameter values or in the carrying capacities. However, for low values of the above-mentioned variables, this type of network seems to stabilize the irregular movements that emerge from a parallel competing/prey-predator network.

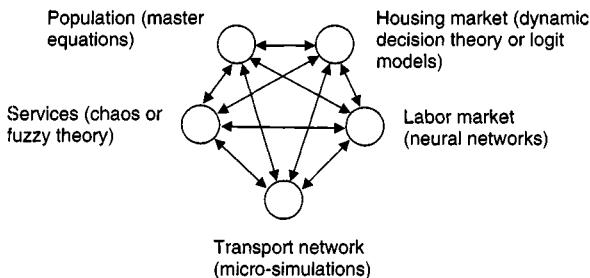


Figure 2. Complex models for complex networks. (Adapted from Haag, 2000.)

- An unstable “corridor” or system in a network can also destabilize the whole network, by increasing the parameter values and carrying capacities of the model.
- By further increasing the complexity of the network by adding multi-layer/multi-level configurations in the above network model we reach the stage of mathematical undecidability, where no suitable information can be extracted.

## 5.2. Complex models for complex networks

The above-mentioned analysis indicates that, if we increase the complexity of the theoretical structure of our niche model in equation (8), e.g. by adding a noise term and thus formulating stochastic differential niche models, we will probably end up with more computational problems. In other words, it seems that only by reducing analytical complexity is it possible to have stable information in the niche models, and thus useful models in forecasting.

An alternative approach to introducing stochasticity and non-linearity in the population models is the adoption of master equations, which are equations of motion for the temporal evolution of the probability distribution of the users. An example of how to integrate master equations in a dynamic traffic flow model is given by Haag (2000). Another way of introducing complexity in the model framework is the development of different sub-models with interfaces. A possible configuration is illustrated in Figure 2.

This would appear to be a fruitful research line, especially in the light of the exploration of the relationships between the micro-behavior and macro-dynamics of population. However, this requires an integration from the theoretical, methodological, and empirical viewpoints of different modeling concepts and frameworks, and, consequently, the building of networks of interdisciplinary, interested researchers and related interfaces.

### 5.3. Detecting complexity from data

A way of detecting complex behavior is the use of techniques able to extrapolate, from data, non-linear network interactions. It should be noted that in chaos and complexity analysis 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. Useful in this respect are, for example, the Brock–Dechert–Scheinkman statistic (Brock et al., 1987) or the method of the largest Lyapunov exponent (Wolf et al., 1985), which have been designed to detect chaos in time series. However, Brock et al. (1987) underlined the ambiguity of the results, by claiming that the models did not show parameter values that allowed for chaotic regimes. Koller and Fischer (2002) carried out a comparative analysis seeking non-linear dependence in time series. The above-mentioned techniques could certainly be used in a complexity framework, even though past experience teaches us that a non-linear structure may be undetectable for several reasons; for instance, because of the interactions between negative- and positive-feedback loops and the inherent noise due to data measurement errors.

An alternative approach for detecting complexity from data analysis is the artificial neural network (ANN) tool, which is part of the bio-computing modeling approach. Insights on the functioning of the human brain and nervous system led to the study of ANNs, where calculation is based on the principle of the distribution of activity to a high number of simple calculation units (neurons), strictly related and working in parallel. In addition to their use for biological metaphors, ANNs are increasingly used as a non-linear statistical adjustment and goodness-of-fit technique for large data sets. Their application appears important in the presence of large, complex networks. In transportation analysis, ANN experiments have been carried out in relation to traffic control, as well as to modal split problems (Hensher and Ton, 2000; Nijkamp et al., 2003). Although in ANNs no modeling hypothesis is needed and no exact function underlying the variables and the data is imposed, ANNs have been related to more conventional statistical methods, as previously outlined. In this context, it is interesting to mention recent work addressing the compatibility between ANNs and MNL/SIMs (Schintler and Olurotimi, 1998; Fischer and Reismann, 2002).

## 6. Network resilience

The issue of the relationship between stability and complexity is still open. From the analytical viewpoint, the formal limit of determining stability solutions for a complex network with a large number of dimensions, parameters, and dynamic flows has still not been resolved. One way to look at this problem is to focus on the

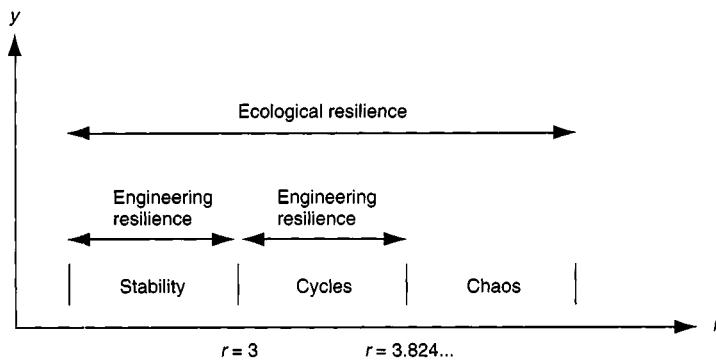


Figure 3. Engineering and ecological resilience for the logistic equation in discrete terms. (Source: Reggiani et al., 2002.)

sensitivity of the network to perturbations, i.e. on the structural stability of the network. Using this approach some typologies of stability can be found – instead of strict stability solutions – for the difference/differential equations governing the network. An example is the results of simulation experiments referring to a network of interrelated logistics niches. New concepts such as fragility and robustness emerged. In general, a dynamic fragile system can be defined as a system that will persist only for tightly circumscribed values of parameters and that will tend to collapse under perturbations to parameters or population values. In contrast, a dynamic robust system will persist for high values of parameters. 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 perturbation of some state variable from a given value” (Perrings, 1994). In particular, two definitions are usually given. Pimm (1984) takes the resilience of a system to be a measure of the speed of its return to equilibrium, while Holling (1973) regards resilience as the perturbation that can be absorbed before the system converges toward another equilibrium state. The first definition is more traditional, focusing on the properties of the system/network near some equilibrium point, while the second is more innovative, looking at the properties of the system/network further away from the stable state, i.e. considering the size of the stability domain. Pimm’s measurement is easier from an empirical viewpoint than that of Holling, however the latter definition may be more interesting by assuming that different states of a system/network involve different equilibria, and that evolution is formed by the switch of these systems/networks from one equilibrium state to another one.

Some authors also call the first concept “engineering resilience,” and the second “ecological resilience” (Reggiani et al., 2002). Consider the logistic equation (7): we can identify engineering resilience not only in the conventional stability period, but also in the cycle period, since here the system, after perturbation, will return in the short or long run, to the limit cycle attractor (Figure 3). Ecological resilience may also embed the chaos period, since here instabilities can flip the system into another behavior regime. In a transport network, resilience can then be associated with a stable situation as well as one of congestion.

Resilience reveals a framework that goes beyond the usual stability concept, since, in principle, a complex system can be resilient. However, to the author’s knowledge, there are no empirical applications of ecological resilience to real cases in the economic–spatial field. One simple way would be to detect resilient behavior from data by means of stability methods, e.g. Lyapunov exponents. However, this empirical solution is not satisfying, showing again the difficulties in implementing complex measures other than the usual stability rules.

More insight into the complexity issue can be gained by integrating parallel or alternative approaches. For example, it may be worth exploring the connections between the resilience concept and other concepts in spatial economics, such as adaptive behavior, learning behavior, path dependence, persistence, survival of the fittest, and small-world networks (Batten, 2000; Wilson, 2000).

## 7. Emergence and self-organized criticality

### 7.1. *The concepts of emergence and self-organization*

Recent debate on complexity has mainly focused on dynamic or evolutionary complexity. In particular, the characteristic of complex networks of high levels of interdependence through non-linearities drew attention to the fundamental feature that the outcome is not obvious from the simple building blocks. On this basis, the concept of emergence came to the fore, by highlighting that the interacting components of a network give rise to global properties that are more than the sum of the parts.

The emergence concept is closely related to the concept of self-organization. Self-organization can be considered as the ability of certain equilibrium systems to develop structures and patterns in the absence of control or manipulation by an external agent (Nicolis and Prigogine, 1977; Jensen, 1998). We can therefore identify self-organization (strictly connected to the synergy concept) as the capacity of a complex system to redefine and develop organized structures, even in conditions far from equilibrium. The emergent phenomenon is then a collective behavior, a self-organized structure, the result of the continuous dynamic interplay

between the macro- and micro-elements of a system or network. Emergence tells us that “an economic system of interacting agents (like bar attendees, 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 themselves. These statistical regularities are large-scale features that emerge purely from microdynamics. They signify order despite change” (Batten, 2000). If we consider the logistic niche, as in equation (8) and Figure 1, the envelopes  $y_1, y_2$ , and  $y'_1$  represent the emergent meso-structures resulting from the interaction with the dynamic behavior of MNL type at the micro-level.

Finally, the emergence concept emphasizes the evolutionary form of organized complexity, in contrast to the disorganized complexity advocated by many researchers in the last century. This paradigm opens the perspective of order in complexity, by reinforcing the idea that complexity is not only chaos and unpredictable behavior. In line with this is Kauffman’s (1993) definition of complex regime as the transition region, on the edge between order and chaos. In addition, he regards the complex regime as the natural culmination or attractor of selective evolution, by arguing that “selection achieves complex systems capable of adaptation. Moreover, there are general principles characterizing complex systems able to adapt. They achieve a poised state near the boundary between order and chaos, a state that optimizes the complexity task the system can perform and simultaneously optimizes evolvability.” This boundary state, to which complex systems seem to tend, e.g. in a traffic jam, is also called self-organized criticality (SOC).

## 7.2. *The concept of SOC*

SOC expresses the criticality behavior of the edge of chaos, i.e. the propagation, through the entire system, of a local distortion. Bak and Chen (1991) define this natural/critical state as follows: “Many composite systems naturally evolve to a critical state in which a minor event starts a chain reaction that can affect any number of elements in the system.” They studied the sand-pile model, by observing, at the critical state, a chain of many tiny and a few large avalanches. They also demonstrated the interesting property of a power law distribution relating the frequency and size of these avalanches (Jensen, 1998). A further interesting property of SOC is the fact that the external driving of the system needs to be much slower than the internal relaxation process, by going through an evolutionary process with meta-stable states and threshold dynamics. SOC may then explain the dynamics of catastrophes, such as earthquakes, forest fires, volcanic eruptions, up- and down-swings in economic markets, political transformations, and spatial phenomena where small shocks can induce the

system to start a chain reaction – an obvious example in a transport network is the traffic jam, where a small shock can provoke catastrophic chain reactions.

By considering the phase-space analysis of the dynamic logistic equation (7) in discrete terms, it is possible to identify the existence of SOC in the boundary area between the ordered and chaotic period (Figure 3). The external driving forces can then lead the system (in a short or long time period) to meta-states, and hence to the edge of chaos. If we then consider dynamic interrelated logistic niches as in equation (8), the carrying capacities  $K_i$  and the slopes of the curves that represent the emergent structures from micro-logistic interacting behavior are crucial for the SOC configuration. It should be noted here that again the logistic niche network of form (8) appears to be a convenient tool for formally interpreting a blend of approaches and concepts, from chaos to complexity/resilience, to SOC.

There is still much debate over the lack of a precise definition and formalism concerning SOC. In particular, it is questioned whether SOC is a unifying theme or just a special case of more general turbulence phenomena, and the need for more thorough studies and empirical applications is stressed. Even though SOC applications are still rare, the experiments carried out so far on transport and spatial systems show interesting results (Reggiani and Nijkamp, 2003). An alternative way of detecting the occurrence of SOC is from data analyses, by testing the power law distribution of the avalanches. It could, therefore, also be interesting to explore this latter methodology in transport networks and telecommunications dynamics.

## 8. Conclusions

The most frequently used dynamic approaches in spatial economics, with reference to their interpretation and applicability in transport, have been revisited in this chapter. The following general conclusions can be drawn:

- Conceptually, recent advances in complexity theory show the possibility of new model-building opportunities and more effective interdisciplinary links; in particular, the new concepts of emergence and SOC indicate the relevance of the impact of micro-behavior on the dynamic functioning of the whole spatial system.
- Methodologically, the logistic niche model appears to be a significant frame of reference for the evolutionary approaches.
- Empirically, the lack of dynamic data often hampers the verification analysis, i.e. to test the speed of change of the parameter values, and hence their correspondence with the complex non-linear patterns incorporated in the models adopted. The approach using statistical tools, which can

identify, from data, non-linear or complex dynamics, seems more satisfying, even though there is still a need for further work, mostly in the field of space-time data analysis.

In conclusion, a blend of advanced approaches belonging to complex systems theory and simulation techniques, together with the availability of rich data sets, seems to be crucial for a good understanding of the mechanisms underlying the complex dynamics of spatial and transport economic processes.

## Acknowledgments

I wish to thank Peter Nijkamp for the useful suggestions, and Patricia Ellman for carefully checking the English.

## References

- Anas, A. (1983) "Discrete choice theory, information theory and the multinomial logit and gravity models," *Transportation Research B*, 17:13–23.
- Bak, P. and K. Chen (1991) "Self-organised criticality," *Scientific American*, Jan.:26–33.
- Batten, D. (2000) "Complex landscapes of spatial interaction," in: A. Reggiani, ed., *Spatial economic science: new frontiers in theory and methodology*. Berlin: Springer-Verlag.
- Batten, D. and D. Boyce (1986) "Spatial interaction, transportation, and interregional commodity flow models," in: P. Nijkamp, ed., *Handbook of regional and urban economics*, Vol. I. Amsterdam: North-Holland.
- Batten, D., J. Casti and B. Johansson, eds (1987) *Economic evolution and structural adjustment*. Berlin: Springer-Verlag.
- Ben-Akiva, M. and S.R. Lerman (1985) *Discrete choice analysis. Theory and application to travel demand*. Cambridge: MIT Press.
- Brock, W.A., W.D. Dechert and J.A. Scheinkman (1987) *A test for the independence based on the correlation dimension*, SSRI Working Paper 8702. Madison: Department of Economics, University of Wisconsin.
- Casti, J. (1979) *Connectivity, complexity and catastrophe in large scale systems*. Chichester: Wiley.
- Domencich, T.A. and D. McFadden (1975) *Urban travel demand: a behavioural analysis*. Amsterdam: North-Holland.
- Fischer, M.M. and M. Reismann (2002) "Evaluating neural spatial modelling by bootstrapping," *Networks and Spatial Economics*, 2:255–268.
- Haag, G. (2000) "New frontiers concepts in spatial and social sciences: towards nested theories," in: A. Reggiani, ed., *Spatial economic science: new frontiers in theory and methodology*. Berlin: Springer-Verlag.
- Hensher, D.A. and T.T. Ton (2000) "A comparison of the predictive potential of artificial neural networks and nested logit models for commuter mode choice," *Transportation Research E*, 36:55–172.
- Holling, C.S. (1973) "Resilience and stability of ecological systems," *Annual Review of Ecological Systems*, 4:1–24.
- Horgan, P. (1995) "From complexity to perplexity," *Scienze*, 324:80–85 [In Italian].
- Jensen, H.J. (1998) *Self-organised criticality*. Cambridge: Cambridge University Press.
- Kauffman, S.A. (1993) *The origins of order*. New York: Oxford University Press.
- Koller, W. and M. Fischer (2002) "Testing for non-linear dependence in univariate time series: an empirical investigation of the Austrian unemployment rate," *Networks and Spatial Economics*, 2:191–209.

- Li, T.Y. and J.A. Yorke (1975) "Period three implies chaos," *American Mathematical Monthly*, 82:985–992.
- Mannermaa, M. (1995) "Alternative future perspectives on sustainability, coherence and chaos," *Journal of Contingencies and Crisis Management*, 3:27–34.
- Martellato, D. (1998) "Innovation, agglomeration and complexity in urban systems," in: C.S. Bertuglia, G. Bianchi and A. Mela, eds, *The city and its science*. Berlin: Physica-Verlag.
- May, R. (1976) "Simple mathematical models with very complicated dynamics," *Nature*, 261:459–467.
- Nicolis, G. and I. Prigogine (1977) *Self-organisation in non equilibrium systems*. New York: Wiley.
- Nijkamp, P. (1975) "Reflections on gravity and entropy models," *Regional Science and Urban Economics*, 5:203–225.
- Nijkamp, P. (1995) "From missing networks to interoperable networks," *Transport Policy*, 2:159–167.
- Nijkamp, P. and A. Reggiani (1992) *Interaction, evolution and chaos in space*. Berlin: Springer-Verlag.
- Nijkamp, P. and A. Reggiani (1998) *The economics of complex spatial systems*. Amsterdam: Elsevier.
- Nijkamp, P., A. Reggiani and W.-F. Tsang (2003) "Comparative modelling of interregional transport flows," *European Journal of Operational Research*, 155:584–602.
- Ortúzar, J. de D. and L.G. Willumsen (1994) *Modelling transport*. Chichester: Wiley.
- Perrings, C. (1994) "Ecological resilience in the sustainability of economic development," in: *International Symposium on Models of Sustainable Development*, Vol. II. Paris.
- Pimm, S.L. (1984) "The complexity and stability of ecosystems," *Nature*, 307:321–326.
- Reggiani, A. and P. Nijkamp (2003) "The fall of the Iron Curtain and the evolution of regional labour markets: a self-organised criticality perspective," in: S. Chang and Y. Okuyama, eds, *Spatial economic modeling of disasters*. Berlin: Springer-Verlag.
- Reggiani, A., T. de Graaff and P. Nijkamp (2002) "Resilience: an evolutionary approach to spatial economic systems," *Networks and Spatial Economics*, 2:211–229.
- Sen, A. and T.E. Smith (1995) *Gravity models of spatial interaction behavior*. Berlin: Springer-Verlag.
- Schintler, L.A. and O. Olurotimi (1998) "Neural networks as adaptive logit models," in: V. Himanen, P. Nijkamp, and A. Reggiani, eds, *Neural networks in transport applications*. Aldershot: Ashgate.
- Wilson, A. (1970) *Entropy in urban and regional modelling*. London: Pion.
- Wilson, A. (2000) "Spatial modelling: conceptual, mathematical and computational challenges," in: A. Reggiani, ed., *Spatial economic science: new frontiers in theory and methodology*. Berlin: Springer-Verlag.
- Wolf, A., J. Swift, H. Swinney and J. Vastano (1985) "Determination of Lyapunov exponents from time-series," *Physica D*, 16:285–317.

*Chapter 15*

## TRANSPORTATION AND URBAN COMPACTNESS

HARRY W. RICHARDSON

*University of Southern California, Los Angeles, CA*

CHANG-HEE CHRISTINE BAE

*University of Washington, Seattle, WA*

### 1. Introduction

Cities vary in their degree of compactness, usually measured in terms of population densities (although the radius of the urbanized area might be a useful alternative measure). Other somewhat more complex possibilities include the concept of median radial distance (Prud'homme and Nicot, 2003), which measures the distance beyond which one-half of the metropolitan population lives, and the index of compactness (Bertaud and Malpezzi, 1998), which measures the deviation in the actual spatial distribution of the population from a cylinder where the height of the cylinder is proportional to the assumed uniform population density of the metropolitan area. Compactness is measured by the extent to which the integration of the actual population density approximates the volume of the cylinder.

This chapter explores the relationships between transportation and the degree of urban compactness. There are several major issues. Higher densities may help to reduce automobile dependence by facilitating shifts to other modes (e.g. transit, bicycling, or walking). On the other hand, to the extent that motorized modes dominate, higher densities mean more congestion and slower travel speeds. Another question is the scope for promoting compactness, especially in the USA, by planning interventions rather than by the pressures of rising land prices. There is a disconnect between the increasing emphasis on policies to make metropolitan areas denser and the overwhelming empirical evidence that most US metropolitan areas are becoming less dense (Fulton et al., 2001). The experiences of Western Europe and many other parts of the world are similar.

## **2. Implications of urban economic theory**

Urban economic theory has dealt with transportation in a very simple way, primarily to facilitate mathematical representation. The standard assumptions are: centralization of jobs in a central business district, surrounded by a suburban ring with land allocated between housing and roads; one-worker households; and a ubiquitous road system. Given urban population and urban area size (determined by adopting some kind of urban boundary constraint, e.g. equality of urban and rural land rents), it is then possible to calculate the number of commuting trips and the amount of land devoted to roads.

Despite this primitive approach, this line of analysis did yield a general finding of value to the discussion of transport and urban compactness. The market equilibrium solution of the standard model results in too much congestion because commuters do not pay for the social costs of congestion. Imposing a congestion tax (e.g. as cents per kilometer traveled) increases the attraction of closer-in locations, and results in a steeper rent (and density) gradient. As a result, the city becomes more compact.

Beyond this important conclusion, the standard theoretical model falls apart as a representation of contemporary metropolitan life. Once employment subcenters are introduced, commuting is both inwards and outwards, and is very difficult to model. While it is possible to make some progress by introducing the concept of commuting sheds (with the implication that workers commute to the nearest employment center), it does not accurately reflect real-world cross-commuting patterns. The problem becomes even worse when we take account of multiple-worker households, because the workers can then commute in different directions. Modeling these flows requires a discrete origin-destination matrix (typically a simulation approach) rather than a theoretical model dealing with continuous space. Thus, almost the sole contribution of urban economic theory is to establish the direct link (positive correlation) between the price of transportation and the degree of urban compactness.

## **3. Historical evolution**

Examination of the historical evolution of cities shows that there is a close relationship between changes in transportation technology and the geographical size of metropolitan areas (Muller, 1995). Urban history, certainly in the USA, can be divided into four phases:

- 1800–1890, walking and horse-drawn cars;
- 1890–1920, electric streetcars;
- 1920–1945, automobiles as recreation;
- 1945–present, the freeway era.

Each of these phases was associated with increasing radius of the city and a decline in urban compactness. Also, in a limited number of large cities (of which New York and Chicago are prime examples in the USA, as are other world capitals), the development of rail transit had a major impact. Contrary to popular belief, these rail lines were a centrifugal force, reducing transportation costs and increasing the attractiveness of locations close to suburban terminals. Transportation has been a liberating force, permitting a degree of suburbanization that would have been impossible without these technological changes: by the 1990s, suburban areas accounted for 83.7% of metropolitan growth in the USA. The strength of the suburbs has been intensified by the development of “edge cities” that can compete with the central business district in attracting office and retail jobs (Garreau, 1991). The challenge for planners today is whether they can reverse the course of history by inducing enough shift in travel modes – from a 98% reliance on auto trips in terms of motorized modes (excluding school buses) or 86.4% of all trips including non-motorized modes, in the USA (Pucher and Renne, 2003 – based on US 2001 *National Household Travel Survey* data) to change the centrifugal forces that have destroyed the compact city.

#### 4. Interpreting data and the urban scale

One of the major problems in comparing variables such as measures of urban density and transportation is the differences in defining the geographical scope of alternative definitions of metropolitan areas (this problem applies to the discussion of almost all urban issues). This topic could be explored extensively, but for reasons of space one aspect will be mentioned supported by an example. In core areas, there is likely to be more reliance on public transit; in peripheral areas where densities are lower and transit services are sparser, automobiles will be the dominant commuting mode. Metropolitan average densities and travel mode shares can be very misleading. As an example, in Central and Inner London in the UK, 60.3% of commuting trips are by public transit, while in Outer London and the rest of the Greater London region 73.8% of trips are by private vehicles (Giuliano, 1998). Population densities in Central London, despite the magnitude of commercial development there, are more than eight times higher than those in the rest of the metropolitan region (Richardson et al., 2000). Although this bifurcation may be an extreme example, it illustrates the dangers of metropolitan-wide generalization.

This example raises a broader question: the issue of spatial scale. We can conceive of at least three scales: micro, meta, and macro. The micro-scale is at the neighborhood level (e.g. the concepts of “healthy neighborhoods” or “walkable neighborhoods” are currently very popular, especially in the USA). The key idea is that redesign to higher densities at the neighborhood level, even in suburban areas,

can stimulate walking and cycling. While this may be true, it appears to make little difference to automobile dependence and vehicle-kilometers traveled. The meta-scale refers to metropolitan areas or cities. Although there are significant differences in densities and degree of compactness in different parts of the metropolitan area, it is much easier to handle transportation issues at the metropolitan level – e.g. organizing public transit or the highway network, or in the cases of compact locations promoting transport-oriented developments (TODs). The macro-scale is the regional level. In the USA, this level does not exist in practice. Regional planning agencies, for the most part, have no implementation powers, and there are no jurisdictions larger than cities or counties. Of course, there are the US states, but most of these are too large to deal with the interconnections between transportation and urban compactness. Thus, most of the interesting discussions focus on either the metropolitan or the neighborhood scale. The latter is particularly interesting because it is much easier to change urban form at the micro-level and databases in the USA (e.g. Tiger files for urban form, and Census tract and block data for travel behavior and related socio-economic characteristics).

## **5. The dynamics of transportation, land use, and urban compactness**

The literature on transportation and land use has emphasized the interdependence between the two, and it is a two-way street: transportation can influence urban form, and urban form may affect travel behavior. However, the benefits from a land use approach are probably modest. First, the settlement pattern is largely determined, so changes in land use are marginal, although there is some debate about how large that margin may be (e.g. in newly developed suburban areas, revitalized core areas, and infill development). Second, travel behavior may be more susceptible to policy interventions than land use preferences. Third, contrary to common belief, any visit to European or Asian cities confirms that compact cities do not fully mitigate the reliance on automobiles; in fact, automobile dependence is increasing, not declining. This is because for all those who can manage their lives in such an environment without automobiles there are many others who will not. Fourth, land use changes are more costly than changes in transportation choices; although the automobile stock changes slowly, it changes much more quickly than the housing stock. Fifth, there is little evidence that higher densities have much impact on automobile ownership or vehicle-kilometers traveled, although they might encourage additional non-motorized trips (Boarnet and Crane, 2001).

Nevertheless, many urban planners and geographers believe that physical planning policy options can solve many social problems. Among the most popular is that the promotion of compact urban structures will deal with the issue of automobile dependence. As suggested above, living in a compact urban

settlement may result in a higher proportion of local non-motorized trips, but the impact on out-of-neighborhood automobile trips is uncertain and probably small (Boarnet and Crane, 2001).

A more important consideration is the difficulty of effecting major changes in the built environment. In the USA, urban settlement patterns are more or less given. There can be small changes at the micro-level (e.g. infill projects and New Urbanist settlements), but the macro-geographical distribution remains more or less the same. As an example, consider Fairfield Village outside Portland, Oregon, a little known model of new urbanism. It is a small community of 600 rather compact and quite attractive dwellings of different types (single-family homes on small lots, townhomes with accessory units, live-and-work row houses, condominiums, and apartments). However, it is accessible only via the freeway. Another Portland example is Orenco Station, with similar housing-type characteristics. This case is somewhat different, however, because it was built with light-rail access as its primary amenity, and has been promoted as a blueprint for transit-oriented development (Bae, 2002). Yet, only 20% of its residents regularly use the MAX light-rail service for the very simple reason that even if you are traveling to downtown Portland (about 25 km) it takes twice as long by MAX as by the private automobile.

Unless further neighborhood design features are incorporated, compactness in itself is not a solution. Evidence in the USA suggests that the elasticity of automobile ownership with respect to population density is about -0.10 (Ingram and Liu, 1999); in other words, a doubling of population density only results in a 10% decline in auto use. The number of cars per unit area significantly increases, and compact cities become more congested and more polluted. These problems might be minimized by an aggressive pedestrian-only section in the city center combined with peripherally located parking structure or lots, but this approach has made little headway in the USA, although it has been adopted extensively in some parts of Europe. Also, this approach works much better in small or medium-sized towns.

Furthermore, the contrast between sprawling USA and compact Europe and, to a lesser extent, compact Asia is narrowing rather than widening, despite major differences in public policies (Nivola, 1999). Take France, for example. Lifestyle preferences favor a quasi-rural life close to a big city (Prud'homme and Nicot, 2003), not only Paris but also other large cities (e.g. Lyon, Marseilles, and Bordeaux). Because of the limitations of public transit services outside the large urban cores (except for inter-city rail service), this lifestyle can be accommodated only via the private automobile.

There is more evidence of a stronger link in the other direction, from transportation to land use. In particular, investments in new highway infrastructure may promote employment growth nearby – (e.g. the toll roads in Orange County, California (Boarnet and Medda, 2003)). However, this is unlikely to result in more

compactness, rather the opposite, because most new roads are built on the metropolitan fringe. Hence, new highway construction tends to promote more job decentralization. Also, even in these cases, the highways–non-residential land uses link may not be causal. Instead, highways may be an intervening variable. First, there is suburbanization of population. Second, the expansion in population stimulates the need for more roads. Third, employment growth occurs. But this could be a lagged response to the previous population growth as mobile employers seek out the suburban labor rather than the result of new highway construction.

A related issue is whether transit investments might promote more compact urban development. Most, but not all, of the new rail transit developments are radial systems that either terminate or pass through the central business district. Some argue that this might help to promote downtown or inner-city development around rail stations, thereby increasing densities. This might happen on a micro-scale, e.g. moderate-size infill development. Also, although there was a modest revival in both population and employment in a few downtowns among large US metropolitan areas in the 1990s (e.g. Denver and Seattle; Birch, 2002), the more successful were not rail cities. Empirical research by Ihlanfeldt and Bollinger (1997) found little evidence of non-residential land use impacts around Atlanta near rail stations. Furthermore, in a much earlier theoretical analysis, Capozza (1973, 1976) argued that a subway accelerates suburbanization by increasing accessibility to downtown – i.e. declining transportation costs pushed the equilibrium location further out. This may be less true now than then, however, because job decentralization has reduced the impact of accessibility to the central business district as an influence on land values.

## 6. Transit-oriented developments

A major concept linking transportation and urban compactness is TODs. The key idea is promote both residential and commercial development around transit terminals (usually, but not necessarily, rail) to reduce auto dependence and promote mixed land uses in a denser urban environment (Cervero, 1998). There are many successful examples in East Asia and, to a lesser extent, Western Europe, but experience in the USA has been much less favorable. A major obstacle is the very low transit share of personal trips (about 1.6%, according to the US 2001 *National Household Travel Survey*) and the appeal of TOD locations to prior transit riders (more than two-fifths of those choosing to live or work in TODs). But there are other difficulties, such as pre-existing land uses with both redevelopment and large-scale land assembly being hard to implement and the fiscal impacts being problematic (Boarnet and Compin, 1999).

Prospects for TODs in the USA are brighter on greenfield sites with new rail stations. One of the most closely watched examples is Orenco Station outside

Portland, Oregon, 25 km west of downtown on the western corridor of the MAX light rail line to Hillsboro (Bae, 2002).

It is too early to judge whether Orenco Station will be a success. It offers a variety of housing types at relatively high densities (by US standards; 2.7 dwelling units per hectare for single-family homes and 9.1 multiple dwelling units per hectare) in an up-scale neighborhood. However, most of the housing is well to the north of the rail station, nearly 2 km away, and is more oriented to a commercial retail strip than to the station itself. Furthermore, the freeway is less than 4 km away, and travel to downtown Portland is twice as fast by the freeway than MAX light rail. Thus, it is not surprising that a survey of residents found that only one in six used transit more than twice a week. Some commuters walk to the nearby high-tech worksites of Intel, NEC, Fujitsu, and Toshiba. Another problem is the lack of parking near the station (except for a 150-car lot for transit carpoolers), given that most homes are not within convenient walking distance and the infrequency of the feeder bus system from more distant access points. Yet, rail ridership is increasing faster than population growth, and housing close to the stations commands a price premium, so the jury is still out.

## 7. New urbanism

New urbanism is an invention of architects. Even its proponents admit that it is a form of social engineering and spatial determinism – e.g. streetscape, housing design, and other characteristics can affect social behavior. Its major promoters include Calthorpe, Duany, and Plater-Zyberk (Calthorpe, 1993; Duany et al., 2000). The core idea is that the adoption of neo-traditional housing designs (e.g. front porches, zero setbacks, back alleys with garages, and accessory units), higher densities, street design elements, and more public than private spaces will lead to all kinds of beneficial effects. These include more communitarian and civic behavior. Another inference is the promotion of social equity, belied by the higher prices of houses in new urbanist communities (Eppli and Tu, 1999). A different issue is the trade-off between the property value protection of homeowners' association regulations (e.g. restrictions on paint colors, window fixtures, flags, signs, and lawn ornaments) and the limitations they impose on individual liberty. But the most relevant issue related to this topic is the impact on travel behavior.

The experience with new urbanist developments is mixed. First, almost all the successful projects (e.g. Seaside and Celebration, Florida; Kentlands, Maryland; Laguna West, California; North Landing and Issaquah Highlands, Washington) were built on greenfield sites. This meant that the developments have been primarily suburban. Second, as a consequence, the lack of transit services, despite *ex ante* promises, has been notable. Third, the more compact spatial structures, combined with a grid street pattern and the avoidance of cul-de-sacs, has

encouraged more on-site walking and bicycling. Fourth, the lag in commercial development, both as workplaces and shops, means that residents typically have to leave their communities for work and for shopping. As a result, the impact on off-site vehicle-kilometers traveled is negligible (Boarnet and Crane, 2001).

## 8. Neighborhood types

One of the most recent contributions to the compact city debate is by Bagley and Mokhtarian (2002). They accept the argument that higher-density, mixed-use communities can result in fewer vehicle trips and smaller distances. However, they question the causality. Instead, they suggest that it is self-selection bias: individuals and households with predisposing lifestyles, travel habits, and attitudes tend to choose these high-density neighborhoods for better compatibility with their preferences. They test this hypothesis using a structural equation model and employing a large number of socio-economic, lifestyle, and attitudinal variables with five different neighborhoods in the San Francisco Bay Area defined as "traditional" or "suburban" on a continuous disaggregated scale, using principal components analysis. The results broadly support the hypothesis, except for the finding that "suburban" Pleasant Hill is associated with more transit use (simply because the *raison d'être* of this location is that it is a transit-oriented development built around a BART (Bay Area Rapid Transit) rail station). This is an artifact of the sample neighborhoods, and does not undermine the general conclusion: "when attitudinal, lifestyle and sociodemographic variables are accounted for, neighborhood type has little influence on travel behavior" (Bagley and Mokhtarian, 2002). They also refer to a paper by Shiftan and Suhrbier (2000) that suggests that while people like the best of both traditional and suburban environments, the bottom line is that the desire for a large lot, bigger house, and attached garage outweighs the convenience of close-by shops and transit access. Another admitted weakness of the study is its cross-sectional character in light of the obvious fact that attitudes can change over time.

## 9. Intertemporal changes

This leads us again into a brief discussion of dynamics. Hickman and Banister (2002) argue in favor of looking at the relationship between transportation and land use over time. A serious limitation of their analysis is that their intertemporal data span is a mere 3 years. As implied in the earlier discussions, land use changes very slowly, so a 3 year change in, for example, density will make a minimal difference to urban form. Perhaps an argument can be made to use cross-sectional data as a substitute for changes over time, but even in this case the evidence is

unclear. Using data from a suburban county near London in the UK, they found that average commuting distance declined by 10% between 1998 to 2001 in the lowest-density areas (less than one person per hectare) and by 19% in areas within the relatively high density range (20–35 persons per hectare), but hardly changed at all (0–2%) in the other density ranges (1–10, 10–20, and >35 persons per hectare). However, focusing on levels, commuting distance in both 1998 and 2001 declines with increasing density.

Thus, Hickman and Banister have a point, but they do not have a strong enough database to make the argument. A somewhat better case in a very different part of the world (Seoul, Korea) is made by Bae and Jun (2002). In Seoul, there is a striking difference in densities between the central city of Seoul and the rest of the Seoul metropolitan area: in Seoul, densities averaged 16 320 persons/km<sup>2</sup> in 2000, 15.8 times more dense than the suburban areas (1033 persons/km<sup>2</sup>). Over a 15 year period (1980–1995), commuting distances increased more over this period for workers who moved out of the city (14.3%) than for those who stayed in the city (10.2%). The more striking statistic, however, is that the low-density suburbanites had a 67.7% longer commute than the high-density city resident-workers (21.90 km rather than 13.06 km). Also, the Seoul metropolitan area data are not distorted by skewed mode choices because the balance among auto, rail, subway, bus, and, even, walking is more evenly distributed than in most other world cities.

Yet another example is the research on Orange County, California (Boarnet and Medda, 2003), mentioned above. It suggests that transportation investments can influence urban form by stimulating both residential and non-residential development in the vicinity of a new corridor. By implication, the reverse is less likely, i.e. changes in urban form affecting transportation. Some planners believe otherwise, but with little supportive evidence.

## 10. Dispersal and travel behavior

A counterpoint to the assumed link between travel and urban compactness is the argument that metropolitan dispersal may be associated with shorter commuting times. If more and more jobs are decentralized, workers may reduce their commutes by moving to the suburbs and avoiding core city congestion. An alternative result is that the commute could become longer if workers are not interested in moving toward journey-to-work minimization and because of the rapid growth of multiple-worker households, not all of whom may be able to find work close to home.

Using panel data for the 1985–1997 period from the American Housing Survey, Crane and Chatman (2003) find some provisional answers to this question. On the whole, more suburbanized employment is associated with shorter commutes, except for manufacturing and financial institutions (perhaps because agglomeration economies and zoning leads to more clustering than dispersal, even in the suburbs).

A different hypothesis (Crane, 1996) is that some households, facing uncertainty about jobs in the future and high moving costs, may tolerate somewhat longer commutes. For example, a worker moving to a new job in the suburbs may continue to live in the central city as a risk-averse strategy. This hypothesis has not been fully tested. If this type of behavior were widespread, it might imply some inertia in travel behavior adjustments as jobs continued to decentralize.

## **11. Information technology**

Another major complication potentially affecting urban form is the changes in information technology. These are related to transportation because telecommunications can be substituted for trips, although they may be complements, e.g. telecommuters avoid the commute but tend to make more non-work trips. There are individual examples of transportation savings (e.g. remote medical consultations, and arraignments of prosecutions), but their scale and their overall impact are unclear. Conceptually, information technology should be able to break the link between transportation and urban form by making further dispersal efficient. The USA experienced significant economic growth in the 1990s in selected non-metropolitan (i.e. rural areas). Much of this growth (e.g. back-office bank functions) was related to information technology. But rural and small town life is associated with more vehicle-kilometers traveled by automobile because the alternative transportation services are absent. Thus, this type of economic decentralization may be efficient for corporations (e.g. lower labor costs) but may not generate any transportation savings, especially for households.

## **12. International comparisons**

It is well known that cities outside the USA, especially in Asia, but even in Europe, are much more compact (Newman and Kenworthy, 1999). They also have much higher public transit and non-motorized mode shares, so the inference is that automobile dependence and density are strongly and negatively correlated. There is nothing wrong with this empirical observation, but its interpretation has to be handled with some caution.

First, there is a difference between levels and trends. The negative relationship between automobile use and compactness is much more convincing in cross-sectional terms. But the rates of growth in automobile ownership in Europe and Asia are much faster than in the USA, typically twice as fast but often more than that (e.g. in Japan). The differential is much higher than can be explained by the acceleration of decentralization trends in these countries, so clearly there are other forces at work besides urban form. For example, per capita income growth

rates have also been higher in many countries than in the USA, and there is a perennial debate about the relative importance of income and urban form as determinants of automobile use (Ingram and Liu, 1999).

Second, the relative price of transportation modes has to be taken into account, and gasoline prices in the USA are typically about one-third of those in Newman and Kenworthy's comparison cities.

Third, choosing rail as a mode is a function of the geographical coverage of the regional transit system, and many cities outside the USA have had large rail systems in place for many years. Most US cities either have truncated rail systems (e.g. one or two corridors) or no rail at all. The consequences are much more a reflection of public investment policy than urban form.

Newman and Kenworthy (1999) also make a major point about shorter commuting distances in cities outside the USA (approximately 8 km in Asia, 10 km in Europe, and 13 km in Australia compared with 15 km in the USA). However, distance traveled is far less important than travel time. Travel times are comparable in both compact and dispersed cities, as a result of average travel speeds being much faster in dispersed cities because of less congestion and more reliance on the faster modes. Thus, the length-of-commute advantages of the more compact cities outside the USA are eroded by their slower travel speeds. For example, comparing the UK's dense London metropolitan area with the more dispersed metropolitan areas of the USA, round-trip average commuting times are 54.4 min and 58.4 min, respectively, while average travel speeds are 24.8 km and 47.5 km, respectively (Giuliano and Narayan, 2002).

### 13. Conclusions

The evidence is far from conclusive. If we focus on the world's largest cities, especially outside the USA, they usually have a relatively compact core city structure and a long-established transit system (usually rail) in place. So, it is easy to draw the inference that compactness induces public transit use. But decentralization trends remain very strong, and when we look at suburban development trends we find that it often becomes very difficult for transit investments to keep up. Once we begin to look at smaller cities and towns, where economies of scale in public transit are absent, the situation is very different. The choices come down to automobile use, infrequent bus service, or non-motorized modes. The first of these becomes the obvious choice for all but a few. In that case, faster suburban travel speeds compensate for the shorter trip lengths observed in compact cities. Also, in many of the world's cities, automobile use is increasing even in central cities, with the result that compactness tends to be more closely associated with the negative externalities of traffic congestion and poor air quality than in more decentralized metropolitan areas. Compactness will fail as an urban

design and planning strategy unless policy-makers can implement effective policies to reduce automobile dependence. Whether this is within the realm of possibility is unclear.

## References

- Bae, C.-H.C. (2002) "Orenco Station, Portland, Oregon: a successful transit oriented development experiment?" *Transportation Quarterly*, 56:9–15.
- Bae, C.-H.C. and M.-J. Jun (2002) *The determinants of suburbanization, intra-metropolitan migration and commuting in the Seoul Metropolitan Area*. Seattle: Department of Urban Design and Planning, University of Washington.
- Bagley, M.N. and P.L. Mokhtarian (2002) "The impact of residential neighborhood type on travel behavior: a structural equations modeling approach," *The Annals of Regional Science*, 36:279–297.
- Bertaud, A. (2003) "Clearing the air in Atlanta: transit and smart growth or conventional economics?" in: C.-H.C. Bae and H.W. Richardson, eds, *Sprawl in Western Europe and the United States*. London: Ashgate.
- Bertaud, A. and S. Malpezzi (1998) *The spatial distribution of population in 35 world cities: the role of markets, planning and topography*, Working Paper. Seattle: World Bank and Center for Urban Land Economics Research, University of Wisconsin.
- Birch, E. (2002) "Having a longer view on downtown housing," *Journal of the American Planning Association*, winter.
- Boarnet, M. and N. Compin (1999) "Transit-oriented development in San Diego County: the incremental implementation of a planning idea," *Journal of the American Planning Association*, 65:80–95.
- Boarnet, M. and R. Crane (2001) *Travel by design: the influence of urban form on travel*. New York: Oxford University Press.
- Boarnet, M. and F. Medda (2003) "Urban design and travel behavior," in: *STELLA Focus Group 3, Second Meeting*. Arlington: National Science Foundation.
- Calthorpe, P. (1993) *The next new American dream*. Princeton: Principia Press.
- Capozza, D. (1973) "Subways and land use," *Environment and Planning*, 5:555–576.
- Capozza, D. (1976) "Land use in a city with two transport modes," *Southern Economic Journal*, 42:442–450.
- Cervero, R. (1998) *The transit metropolis: a global inquiry*. Washington, DC: Island Press.
- Crane, R. (1996) "The influence of uncertain job location on urban form and the journey to work," *Journal of Urban Economics*, 39:342–356.
- Crane, R. and D. Chatman (2003) "Traffic and sprawl: US commuting evidence for 1985 to 1997," in: C.-H.C. Bae and H.W. Richardson, eds, *Sprawl in Western Europe and the United States*. London: Ashgate.
- Duany, A., E. Plater-Zyberg and J. Speck (2000) *Suburban nation*. New York: North Point Press.
- Eppli, M.J. and C.C. Tu (1999) *Valuing the new urbanism: the impact of the new urbanism on prices of single-family homes*. Washington, DC: Urban Land Institute.
- Fulton, W., R. Pendall, M. Nguyen and A. Harrison (2001) *Who sprawls the most? How growth patterns differ across the U.S.* Washington, DC: Brookings Institution (<http://www.brook.edu/es/urban/publications/fulton.pdf>).
- Garreau, J. (1991) *Edge city: life on the new frontier*. Garden City: Doubleday.
- Giuliano, G. (1998) "Urban travel patterns," in: B. Hoyle and R. Knowles, eds, *Modern transportation geography*, 2nd edn. Chichester: Wiley.
- Giuliano, G. and D. Narayan (2002) "Another look at travel patterns and urban form: The US and Great Britain," *Urban Studies*, forthcoming.
- Hickman, R. and D. Banister (2002) "Reducing travel by design: what happens over time?" in: *5th Symposium of the International Urban Planning and Environmental Association*. Oxford.
- Ihlanfeldt, K.R. and C. Bollinger (1997) "The impact of rapid rail transit on economic development: the case of Atlanta's MARTA," *Journal of Urban Economics*, 42:179–204.

- Ingram, G.K. and Z. Liu (1999) "Determinants of motorization and road provision," in: J.A. Gomez-Ibanez, W.B. Tye and C. Winston, eds, *Essays in transportation economics and policy: a handbook in honor of John R. Meyer*. Washington, DC: Brookings Institution.
- Muller, P.O. (1995) "Transportation and urban form: stages in the spatial evolution of the American metropolis," in: S. Hanson, ed., *The geography of urban transportation*, 2nd edn. New York: Guilford Press.
- Newman, P. and J. Kenworthy (1999) *Sustainability and cities: overcoming automobile dependence*. Washington, DC: Island Press.
- Nivola, P.S. (1999) *Laws of the landscape: how policies shape cities in Europe and America*. Washington, DC: Brookings Institution.
- Prud'homme, R. and B.-H. Nicot (2003) "Urban sprawl in France in recent decades," in: C.-H.C. Bae and H.W. Richardson, eds, *Sprawl in Western Europe and the United States*. London: Ashgate.
- Pucher, J. and J.L. Renne (2003) "Socioeconomics of urban travel: evidence from the 2001 NHTS," *Transportation Quarterly*, 57:49–77.
- Richardson, H.W., C.-H.C. Bae and M. Baxamusa (2000) "The compact city in developing countries," in: M. Jenks and R. Burgess, eds, *Compact cities: sustainable urban forms for developing countries*. London: Spon.
- Shiftan, Y. and J. Suurbier (2000) "The effects of land use policies on regional travel using a residential choice stated-preference model and an activity-based travel model," in: *2000 International Association for Travel Behavior Research*. Gold Coast.

*Chapter 16*

## COMPUTABLE GENERAL EQUILIBRIUM ANALYSIS IN TRANSPORTATION ECONOMICS

JOHANNES BRÖCKER

*University of Kiel*

### 1. Introduction

Computable general equilibrium (CGE) analysis, pioneered by Johansen (1960), Harberger (1962), and, on a larger scale, by Shoven and Whalley (1984), is now a standard tool in empirical economics for simulating the effects of variations in exogenous variables and parameters on any kind of economic variable such as output, employment, prices, income, and welfare. Exogenous variations (also called shocks) range from policy variables such as tax rates, tariffs, and transfers over regulatory frameworks to technologies and preferences. The literature is huge. A well-known large multi-country world model is GTAP (Hertel, 1997). A more recent field of application is transport economics, the subject of this chapter. A typical transport economics application is the study of the quantitative impacts of transport initiatives such as infrastructure investments or pricing policies on economic variables.

The basic idea of CGE analysis (also called applied general equilibrium (AGE) analysis) is to take a textbook general economic equilibrium model and to “fill it with numbers.” Its theoretical background is the Walras–Arrow–Debreu theory of general equilibrium, with modern modifications and extensions allowing for imperfect markets, a public sector, money, externalities, and other complications disregarded in the original theory. The approach is called computable (or computational) because it is made for numerical calculations on the computer, unlike the original theory, that was invented just for understanding how prices and quantities are determined in interdependent markets, not for empirical application. The approach is called equilibrium analysis, because its kernel is the concept of market equilibrium as used in micro-economics. It means that all agents in the economy make mutually consistent plans, such that no agent (no firm, no household, no public institution) has an incentive to revise his or her plan. The approach is called general (in contrast to partial) because all market interactions are taken account of. This means two things. First, there are neither “Black holes” for payments to vanish in, nor mysterious fountains spitting money that agents receive.

Any payment of an agent in the model (e.g. transport costs) is a receipt of another agent in the model (e.g. a firm producing logistic services); and any receipt of an agent in the model is a payment of another agent in the model. Second, all agents balance their budget; they expend exactly what they obtain. This does not preclude debt and credit, of course; but any debt and credit must be explicitly handled in the model, such as sales and purchases of goods, and services and factors of production.

CGE analysis is particularly useful in transport economies when traditional cost-benefit (CB) analysis is insufficient for evaluating the welfare effects of transport initiatives comprehensively. The traditional CB evaluation of a new road, say, measures the benefit by the consumer surplus of users generated by reducing generalized costs, and subtracts building costs in market values and the net increase of technological external costs caused by existing and induced traffic. This procedure is fine as long as the following conditions hold: (1) markets are perfectly competitive and cleared by fully flexible prices; (2) welfare distribution is not an issue, i.e. each monetary unit counts equally, irrespective of who gets it; and (3) technological externalities outside the transport sector are negligible.

If one of these conditions is not met, one must extend traditional CB analysis by considering repercussions to the rest of the economy, outside the segment of the transport system directly affected. If only condition (3) is violated, one must look at changes in land use, production, consumption, employment, etc., that may generate technological externalities such as atmospheric pollution or noise. To this end, CGE analysis is one possible analytical tool, but it has competitors such as LUTI models, econometric macro-models, system dynamics models, and possibly more. Here, the debate on the relative merits of CGE and other methods for quantifying indirect effects on a wide range of externality generating variables will not be entered into. But if conditions (1) or (2) are violated, we are in the realm of welfare economics, and there appears to be no viable alternative to computable equilibrium analysis (possibly partial but preferably general equilibrium analysis). Summarizing this, CGE is the method of choice, if the results of conventional CB analysis have to be corrected due to imperfect markets, or if one is not content with aggregated monetary welfare measures and needs distributional details about welfare implications.

Section 2 introduces CGE analysis. Section 3 explains more concretely how the method can be applied to transportation issues. Section 4 discusses the extension of the conventional approach by introducing imperfect markets and dynamics. Section 5 illustrates the applicability of CGE analysis to transport project evaluation by a practical example, and Section 6 concludes the chapter.

## 2. A primer in CGE analysis

As already mentioned, CGE analysis is a numerical computer version of the micro-economic textbook model of general equilibrium. In the textbook model,

agents are individual firms and households characterized by technologies and preferences and interacting on markets. The economy is thought of as a state where all these agents behave rationally, given their technological and budget constraints, and where prices have adjusted such that all markets clear simultaneously.

In the theoretical textbook model one imagines millions or billions of individual agents and billions and trillions of different commodities; one even works with infinitely many agents and infinitely many goods in theory. The forms of technologies or preferences are not very specific; only very general assumptions such as closedness, boundedness, convexity, and monotonicity of production possibility and preference sets are made. One tries to state the weakest possible assumptions allowing the derivation of behavioral functions (supply and demand functions) within the rationality framework, which are both well defined and sufficiently regular to allow a positive answer to the theoretical question in hand (existence of an equilibrium, uniqueness, stability, and so forth).

In empirically applied models, in contrast, there are only a limited number of agents and different goods: tens, hundreds, or perhaps thousands as in recent applications. Even 1000, however, is a tiny number compared with the myriad of agents and of goods in the real world! The forms of preferences and technologies also need to be precisely stated; they have to be numerically specified such that the resulting behavioral functions can be calculated on a computer.

How can one portray a real economy by a simplified model with a limited number of agents and goods? Two ploys help:

- Using representative agents. Instead of trying to model the behavior of all individual firms in an industry of a country or region, for example, one deals with only one single firm representing the entire industry in the country or region (the representative firm). The industry's output is conceived of as the output of one (or many identical) representative firms. In a similar way, the private household sector (or a subsector such as the low-income segment, for example) is portrayed by a representative household.
- Using aggregate goods. Instead of trying to model production and consumption of a huge number of different goods within a certain sector, the entire sectoral output, for example, is conceived of as a single homogeneous aggregate commodity or service (aggregate good). How can one measure the quantity of such an aggregate good representing commodity bundles that are very diversified in the real world? Transport planners measure quantities in terms of tonnes, cubic meters, or similar quantity units. CGE modelers have a different concept: they use artificial units of measurement for quantities, implicitly defined by an arbitrary choice of a price for each aggregate good prevailing in a benchmark equilibrium. For example, let  $y$  denote the output value of an aggregate

good in euros per annum in a benchmark equilibrium and  $p$  the arbitrarily fixed benchmark price of the good, then the quantity is  $x = y/p$  artificial units. To put it a different way: one artificial unit is the quantity costing  $p$  euros in the benchmark situation. An obvious choice is  $p = 1$ , such that benchmark values equal benchmark quantities. In this case quantities are often reported in terms of euros, which is somewhat confusing. What is meant is “benchmark-euro,” which should be understood to be the name of the artificial unit that is chosen such that one unit costs 1 euro in the benchmark equilibrium.

Regarding preferences and technologies, CGE analysis offers a tool kit of specific functional forms containing a number of free parameters allowing the equilibrium solution to be tuned such that empirical observations are reproduced by the equilibrium solution of the model. Fixing the parameters is called “calibration.” After goods and agents have been specified, functional forms chosen, and parameters have been calibrated, exogenous changes to parameters or data are introduced (sometimes called “shocks”), and “counterfactual” equilibrium solutions representing the after-shock world are found.

A CGE analysis can be carried out in nine steps, to be explained next. Consider a very simple example as an illustration. The example is a computable equilibrium analysis, though – for the sake of simplicity – not a general one but a computable partial equilibrium (CPE). This CPE requires all the steps that are necessary in a CGE. Let us assume that a simple aggregate good is produced in a region  $r$ . The output value in a base year equals 1 million euros, say. The good is delivered to region  $s$ , where it is consumed. The total transport costs for transferring the good from  $r$  to  $s$  amount to 0.1 million euros, such that the consumption value in region  $s$  is 1.1 million euros, including transport costs. We expect an improved transport infrastructure to reduce transport costs per unit of commodity to one-half of its benchmark level. (This cost reduction is our “shock.”) The repercussions in sectors of the economy other than the producers in  $r$  and the consumers in  $s$  are assumed to be negligible (hence the partial equilibrium character of the example). The cost saving may benefit producers or consumers, or both. Who will gain, and by how much?

Figure 1 portrays the situation.  $p_0$  and  $q_0$  denote the producers’ cost and the cost to customers before the shock, respectively.  $S_0$  is the supply (equal to demand) before the shock. The difference  $\pi_0 = p_0 - q_0$  is the transport cost per unit shipped. When transport costs are reduced to one-half of their former level,  $p$  rises to  $p_1$ ,  $q$  falls to  $q_1$ , and  $S$  increases to  $S_1$ . Areas A plus B, the consumers’ surplus, measure the welfare gain on the customers’ side, and areas C plus D, the producers’ surplus, the welfare gain on the producers’ side. The total welfare gain slightly exceeds (by the two gray triangles B plus D) the direct cost saving that would be obtained with a constant quantity (areas A plus C). Price and quantity effects as

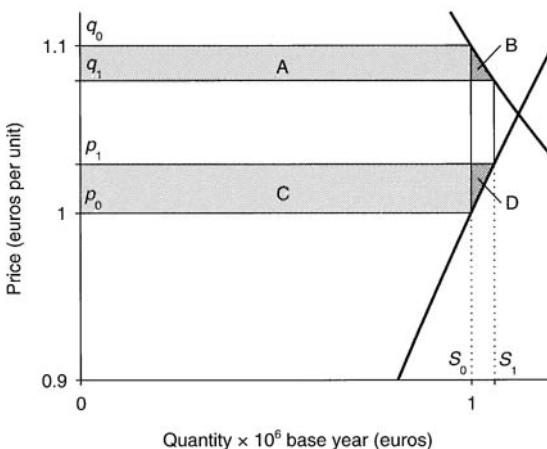


Figure 1. Welfare effects of a transport cost reduction: a PCE.

well as the distribution of welfare gains between producers and customers depend on the respective slopes of the two curves. Steep (flat) curves imply a small (large) increase in quantity. The welfare gain goes mainly to the customers (producers) if the supply curve is flat (steep) in comparison to the demand curve (for more on this, see point 6 below).

Taking this simple example, we demonstrate step by step how to quantify price, quantity, and welfare effects, explaining the general procedure in each step:

- (1) Delineate aggregate goods (in the example: just one good) and specify representative agents (in the example: one producer and one consumer). In a typical CGE analysis we would regard the output of an industry in a region as an aggregate good, for example, such that the number of goods equals the number of industries multiplied by the number of regions. In this case, each industry in each region is portrayed by a representative firm. Other agents are the representative households in each region and the public sector (possibly split up into local and national levels). In a multiregional framework, goods from the same industry but from different producing regions are conceived of as distinct (the so-called Armington (1969) assumption). They may be close substitutes, but they are not perfect substitutes, because this would lead to unrealistic corner solutions: there would be no cross-hauling, only a small number of interregional flows would be positive (as in the solution of the transportation model in linear programming), and an arbitrary small price advantage of a region would attract 100% of demand for the good.

	$I_1$	$I_2$	$L$	$K$	$H_1$	$H_2$	$\Sigma$
$I_1$	1	3			4	9	17
$I_2$	6	1			2	5	14
$L$	7	5					12
$K$	3	5					8
$H_1$			6	0			6
$H_2$			6	8			14
$\Sigma$	17	14	12	8	6	14	

Figure 2. An example SAM.

- (2) Organize benchmark values of transactions between agents in a social accounting matrix (SAM). In the example, benchmark data consist of just two values: the transaction value as the cost to the producer (1 million euros), and as the cost to the customer including transport costs (1.1 million euros). In a SAM, for a full CGE model each agent appears twice, once in a row, and once in a column. The row records the agent's payments, and the column the agent's receipts. General equilibrium consistency requires equality of row and column totals.

A small example SAM is given in Figure 2 for a static economy with two industries ( $I_1$  and  $I_2$ ), two factors (labor  $L$  and capital  $K$ ), and two households ( $H_1$  and  $H_2$ ) (no public sector, no taxes, and no savings or investments).

Industry 1 pays 1 unit for inputs from itself, 6 units for inputs from industry 2, 7 units for labor, and 3 units for capital (similarly for industry 2). Labor income (12 units) goes half and half to the two households, while capital income (8 units) only goes to household 2. Household 1 spends 4 units of income (6 units) for goods from industry 1, and 2 units for goods from industry 2 (similarly for household 2). Gross production is 31 units, of which 11 units are intermediate goods. The GDP is 20 units.

- (3) Choose the market form (in the example: price-taking behavior). Early CGE models relied almost exclusively on the assumption of perfect competition, while recent research also models monopolistic or oligopolistic behavior. The most popular approach in recent research is to assume monopolistic competition with diversified goods. In this case the aggregate good is no longer homogeneous but split up into a large number of symmetrical varieties. Imperfect competition models are more demanding in terms of data requirements and solution techniques. In recent CGE research a new approach originating from Dixit and Stiglitz (1977) allows

Armington's assumption to be dispensed with: the goods (even those within one region and one industry) are conceived of as bundles of a large number of heterogeneous varieties. Each variety is produced exclusively by a single firm. As a consequence, there is a limited degree of substitutability between source regions, because the varieties from different sources are not the same (see also Section 4.1).

- (4) Fix benchmark prices arbitrarily. In the example let us choose  $p_0 = 1$  such that  $x_0$  equals 1 million units ("benchmark-euros"). The subscript 0 denotes benchmark figures. Note that one can fix only one price per aggregate good. With  $p_0 = 1$  the inclusive price  $q_0$  is implicitly fixed at  $q_0 = 1.1$  in our example.
- (5) Choose functional forms describing the agents' behavior. In the example we assume an isoelastic supply,  $S = \alpha p^\mu$ , and demand,  $D = \beta q^{-\nu}$ , where the Greek variables are parameters. These forms are log-linear; they are popular because the parameters  $\mu$  and  $\nu$  controlling the slopes of the curves (the elasticities) are dimensionless. Another obvious choice for a functional form would be a linear form. In a typical CGE one deals with behavioral functions making not a single quantity but a vector of quantities that are dependent on a vector of prices. As there is a potentially infinite number of functional forms, criteria for choice are needed:
  - computational simplicity – it should not take too much computer time to calculate function values;
  - theoretical consistency – the function should fulfil all restrictions implied by the theory of rational behavior of firms or households;
  - flexibility – the function should allow the reproduction of any desired behavior with a sufficient degree of approximation by choosing the parameters appropriately;
  - parsimony – there should not be too many parameters that require calibration.

Obviously, there is a trade-off between the last two criteria. The theory of functional forms offers different (stronger or weaker) operational definitions of these criteria and shows that it is difficult (sometimes impossible) to fulfil them all. Any choice must be a compromise between competing objectives (for more on this see Diewert and Wales, 1987).

- (6) Calibrate parameters. Typically there are two types of parameter, namely "position parameters" (sometimes called "shift parameters" or "shift and share parameters"), shifting supply and demand curves leftwards or rightwards (here:  $\alpha$  and  $\beta$ ); and parameters controlling the slopes (in the example: the elasticities  $\mu$  and  $\nu$ ). Usually one only has a benchmark data set on the value of transactions (the SAM). These allow only for a calibration of the position parameters. They are chosen such that benchmark data are exactly reproduced. In our example we obtain  $\alpha = 1 \times 10^6$  from  $S_0 = \alpha \times 1^\mu = 10^6$  units, and

$\beta = 1.1^\nu \times 10^6$  from  $D_0 = \beta \times 1.1^{-\nu} \times 10^6$  units. Furthermore,  $\pi_0 = 0.1$  euro per unit, such that  $q_0 = p_0 + \pi_0 = 1.1$  euros per unit.

Slope parameters can be obtained from the data only if we observe how the economy reacts to controlled exogenous shocks. As it is fortuitous to have a full SAM for just one benchmark year, we cannot expect to compile a time series of such data. Even if we had such a time series, we would lack sufficient knowledge about the shocks generating the changes over time. Therefore, research on parameter estimation within a general equilibrium framework is still in its infancy.

The solution to this dilemma is guesswork plus gathering results from econometric research that is usually based on partial equilibrium approaches. A wide range of arbitrariness is to be expected, and this is clearly a weak point of CGE modeling, which one tries to mitigate by sensitivity analysis (see point 9 below). Let us in our case just assume that  $\mu = 2$  and  $\nu = 3$ . Note that these assumptions essentially predetermine the result in our example. If  $\mu/\nu$  is large (small), the welfare gain accrues mainly to the consumers (producers).

- (7) Simulate policy effects by solving the model for counterfactual equilibria. Inserting the parameters into the equilibrium condition of the example,  $\alpha p^\mu = \beta(p + \pi)^\nu$ , yields the unique (real) solution  $p_1 = 1.0291$  for  $\pi_1 = 0.05$ . The subscript 1 denotes counterfactual equilibrium values. Hence,  $q_1 = 1.0791$ . The manufacturer's costs rise by about 2.9%, the consumer's costs fall by about 1.9 %, and output rises by about 5.9%.

In our simple example model we only had to solve a single equation for a single unknown, the price  $p$ . Even though this equation is non-linear, this could be done without a computer. However, we usually need to consider large systems with many equations and unknowns, and thus require computationally intensive iterative techniques such as Newton's method or continuous path following to find solutions (Allgower and Georg, 1990). The last two decades have seen enormous advances in computing – cheaper and more powerful hardware and better software – that allow very large systems with thousands of unknowns to be analyzed.

- (8) Calculate welfare gains or losses. In the example we measure welfare effects by producers' and consumers' surpluses (the shaded areas in Figure 1). Integrating the supply and demand curves yields a producers' surplus of 29.995 euros (areas C plus D in Figure 1) and a consumers' surplus of 21.471 euros (areas A plus B in Figure 1). The total surplus is just a little (1.466 euros) more than the direct cost saving of 50 000 euros, which would be obtained by multiplying the cost saving per unit (0.05 euro) by the benchmark quantity (1 million units). In other words, the welfare gain due to "induced" traffic (areas B plus D in Figure 1) is negligible in our example.

In a general equilibrium, welfare is not measured by producers' and consumers' surpluses but directly by translating utility changes of households into monetary terms. Although the producers' surplus goes to the firms supplying the respective commodity, it is not the firms who are the ultimate beneficiaries but the private households, to whom the gain is transferred either by the additional profits of firm owners or by increases in the rewards of the factors employed in the respective firms. In a general equilibrium, these transfers of gains to the ultimate beneficiaries are explicitly modeled, and it is therefore possible to measure the welfare gains where they eventually accrue, namely on the households' side. For each representative household in the model, the welfare gain of a transport cost reduction, say, is the utility increase in the low transport cost equilibrium as compared to the high-transport cost equilibrium. This gain is translated into monetary terms by Hicks' concept of equivalent variation (EV). The EV is defined as the extra income one would have to pay to a household in the high-transport cost equilibrium in order to make it as well off as it actually is in the low-transport cost equilibrium. This gain covers all direct and indirect effects, those resulting from changing factor returns, those due to changing profits of firms owned by the households (if there are any profits), as well as those stemming from price changes for consumed goods and services. If travel time is an argument in the representative households' utility functions, then the monetary value of travel time savings is also covered by the EV measure (Bröcker, 2002).

- (9) Check sensitivity with respect to parameters not derived from the benchmark data (typically elasticities). In our example the results are obviously highly sensitive with respect to the assumed elasticities  $\mu$  and  $\nu$ .

What distinguishes a CGE analysis from our simple example is, first, that we have many goods and agents and therefore many prices and quantities to be determined; the second distinction is that our example is a partial equilibrium model, not a general equilibrium one. In a partial equilibrium we cut out one market (or several markets) without caring about the rest of the economy. We are uninterested in such questions as "Where do households get their income from?" or "Where do firms' returns go?" As already stated in the introduction, loose ends such as these are not allowed in a general equilibrium: each agent's budget constraint has to hold, and each monetary flow appears twice, once as a payment, and once as a receipt.

### 3. Transportation in CGE analysis

CGE models in transportation research can be single-region or multi-region models. In a single-region model the transfer of goods and people from one

location to another cannot be dealt with explicitly. Essentially, the transportation sector is handled just like any other sector in this application type. The sector produces one or more classes of transport services that are used as intermediate inputs by firms or consumed by private and public households. Typically, the production of a transport service is assumed to generate negative externalities that harm households or destroy real assets. For policy experiments one can introduce taxes, tradable permits, and other instruments to internalize externalities in order to study the efficiency and distributional implications of these instruments (Conrad, 1997; Conrad and Heng, 2000; Mayeres, 2001).

More demanding are multi-regional models, where transport appears as an activity transferring goods and people between locations. Multi-regional models are used for many other applications than transport (Trela and Whalley, 1986; Jones and Whalley, 1988; Hirte and Genosko, 1988–1989; Madden, 1992; Gazel, 1996; Ando and Shibata, 1997; Bröcker, 1998a; Haddad and Hewings, 2000a). Multi-regional models for transport project evaluation are relatively recent (Buckley, 1992; Bröcker, 1998b; Venables and Gasiorek, 1998; Ueda et al., 1999; Haddad and Hewings, 2000b). In the following, we first deal with goods transport and then with passenger transport.

### 3.1. Goods transport

Modeling the transport service realistically brings a considerable amount of additional complexity into a CGE system. A technology has to be specified for the transport activity and its parameters calibrated, and each transport activity that produces inputs has to be modeled. Because of this complexity and the lack of data for calibrating such a detailed picture of transport activity, simplified approaches are looked for in CGE applications.

A popular simplification is the so-called iceberg approach, originating (though not under this name) from von Thünen (1826), later introduced by Samuelson (1954), and made popular by Krugman (1991). The simple assumption is that transport partly uses up the transported good itself. In von Thünen's work the cost of transporting wheat is the wheat needed to feed a horse drawing a cart. The simplification came to be termed the “iceberg approach” if the transported goods are assumed to be consumed during transportation – like the melting of an iceberg. If  $x$  is the quantity of goods sent from the origin  $r$ , then  $x/\tau$  with transport cost factor  $\tau > 1$  is the quantity arriving at the destination  $s$ , where  $\tau$  is an increasing function of transport distance, with  $\tau = 1$  for a distance equal to zero. If  $p$  is the cost to the manufacturer in location  $r$ , then  $p\tau$  is the cost to the customer in location  $s$ . Hence, the value of the flow in  $r(px)$  is the same as in  $s$ :  $p\tau x/\tau = px$ . Using up resources is thus modeled in a way that is consistent with the general equilibrium framework, and no explicit transport sector needs to be introduced.

This is a comfortable, though not very plausible, concept. A more realistic approach assumes that there is a special sector, from which one has to buy a certain amount of a transport service in order to transfer goods between two locations. This sector can be located at the origin (as in Buckley's (1992) model),<sup>a</sup> the place of destination<sup>b</sup> or possibly partly in the place of origin and partly in the place of destination. If one is particularly interested in the transportation sector itself, one would like to model this sector in detail, regarding intermediate and factor inputs. If the focus is more on the consequences of transport cost reductions on other parts of the economy, one would be content with a rough modeling, assuming for example that an aggregate output of the destination region is one-to-one transformed into transport service. Results do depend on the amount of transport cost reductions, but not much on the way the transport sector itself is specified.

Like any other technology, transportation has position parameters and possibly elasticities (if the service is produced with several inputs, among which different degrees of substitutability can prevail). The former allows the reproduction of any desired level of transportation costs for each aggregated good and each pair of regions in the benchmark equilibrium. These transportation costs must be known for the benchmark in order to calibrate the technology. Usually this information is not available in total from primary sources in practice. Instead, functional relations between transport costs and distances along shortest routes through a transportation network have to be relied on, with links valued by a weighted sum of length and travel time ("generalized distance," see below). As transport costs increase less than proportionally with distance, a plausible simple approach is to assume that the share of transport costs in the value of goods shipped from region  $r$  to region  $s$ ,  $f_{rs}$ , is  $f_{rs} = \xi g_{rs}^\omega$ , with a generalized distance  $g_{rs}$ , and parameters  $\xi > 0$  and  $0 < \omega < 1$ . Sample data on generalized distances and transport costs would suffice to estimate this equation.

As with any other sector in the economy, the transport sector can be imperfectly competitive. In this case prices may not reflect average and marginal costs. They can contain monopoly mark-ups, and limited arbitrage possibilities might allow for considerable price discrimination. If one has very good information on the industrial organization in the sector, these complications can be introduced into the CGE framework by explicitly modeling imperfect competition in the transport sector. However, the author is unaware of any such attempts.

<sup>a</sup>Buckley (1992) implemented a model with three regions and five sectors for the USA with a fixed ratio of required transport services to the quantity of goods shipped. The transport service comes from the local goods sector in the region of origin.

<sup>b</sup>In Hussain's (1996) model, which is experimental without real data, the transport service arises from the output (there is just one sector per region) in the destination.

### 3.2. Passenger transport

Passenger travel is an activity associated with firms (business travel), private households, and the public sector. We will disregard the last category. Business travel is partly tied to goods transport. The cost of inter-regional trade in goods and, in particular, in services is not the mere cost of goods transportation in the narrow sense of the word but also the cost of passenger transport, in so far as the goods trade is accompanied by the travel of persons transferring information, performing service activities, and marketing the products. Hence, the transport costs referred to in the preceding section partly have to be understood as the costs of personal business travel. In a CGE this is taken account of by applying a wider concept of the transportation activity for goods and services (Knaap and Oosterhaven, 2001): the output of the activity is the transfer of goods from an origin to a destination, including the passenger travel required to undertake the aforementioned complementary services. The activity requires inputs for goods transportation in the narrow sense as well as for business travel. The latter may be a function of passenger travel times and travel distances as well as of other cost components such as fuel taxes. As a consequence, policies only affecting passenger travel have an impact similar to policies affecting goods transport in such a framework.

Not all business travel is related to the inter-regional transfer of goods and services, however. Firms need to travel to acquire technological as well as market information, to establish cooperation and contacts, and for many other purposes. This can be incorporated into the modeling framework by introducing passenger travel as another input of firms, just like the input of goods, factors, and services (Ueda et al., 1999). For some firm in region  $r$  the vector of trips to destinations  $s = 1, \dots, n$  (and to return) is part of the input vector, with each component of the vector having its respective price. Policy experiments change the resources required for these trips and, hence, the respective input prices.

Business travel, however, is only a minor part of passenger travel in the real world. Commuting, leisure, and shopping travel are more significant components. An obvious way of handling leisure or shopping travel in the standard framework is to regard them as a kind of trade in services. Travel of a household residing in  $r$  to a destination  $s$  (and to return), say, is then the same as the household's demand for a service from supply region  $s$ . In addition to the monetary costs of travel, time costs have to be taken into consideration. While the micro-economics of travel demand are well understood, CGE applications that include leisure and shopping travel are in their infancy. The first steps in this direction have been taken in the IASON project (Bröcker et al., 2001; Bröcker 2002). Modeling commuting requires endogeneity of the location of residence or employment, or both. An interesting initial advance in this direction is an urban equilibrium model by Horridge (1994), introducing preferences for the locations of residence and of

work, randomly varying for households. The approach allows the impact of changing commuting costs on the spatial distribution of the population and jobs, on land prices, and on commuting to be simulated.

### *3.3. Economic equilibrium and transport network equilibrium*

An important practical issue in transportation research is how to measure the “generalized distance” – a summary measure reflecting out-of-pocket costs as well as time costs. It is usually obtained from a network with arcs distinguished by mode, type of route, etc., and with weights assigned to each arc category. The weights differ between types of goods, such that different distance measures are obtained for different industries. This is the link between the economic CGE model and the transport planner’s mode choice and route choice model (Friesz et al., 1998). The state of the art is to integrate modal split and assignment in a stochastic user equilibrium model based on a hyper-network (Sheffi, 1985). The arcs of this network are routes for specific modes and of a specific quality as well as transfer links between modes. Perceived distances along arcs randomly vary over users, and increase with the expected loading of the network, producing congestion.

The outcome of such a network model is an expected generalized distance for each origin–destination pair. In a congested network it is itself a function of the volume of flows through the network. This interdependence generates an externality, implying that even a perfect competition equilibrium for the whole system generates an inefficient allocation. Hence, we have the ideal framework to study internalizing congestion externalities by pricing measures. Unfortunately, however, incorporating the full interdependency between transport volumes and congestion costs into a general equilibrium framework is computationally rather demanding for realistic networks.

## **4. Extensions**

### *4.1. Imperfect markets*

The standard Shoven–Whalley approach to CGE analysis assumes perfect competition. Once we leave the narrow world of perfect competition, many other forms of market structure are possible. CGE research in international trade has shown that the estimated welfare impact of transaction cost reductions can change dramatically with a deviation from the perfect competition assumption. The same holds true for dropping the assumption of flexible wages clearing the labor market. If transaction cost reductions lead to employment expansion, there is an extra welfare gain beyond the cost saving itself, because the social marginal cost of

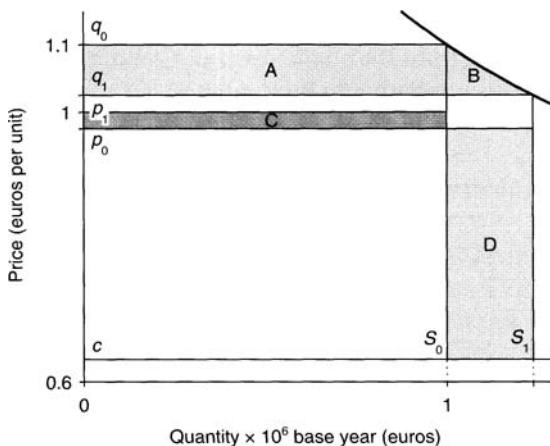


Figure 3. The welfare effects of a transport cost reduction: a CPE with monopoly supply.

employment falls short of the social marginal return reflected by the wage rate. Experimental CGE studies (i.e. computer simulations with small artificial data sets) with imperfect markets have been presented by Hussain (1996), Hussain and Westin (1997), and Venables and Gasiorek (1998).

The importance of the market structure can be learnt from a modification of our basic CPE example in Section 2 (see also Standing Advisory Committee on Trunk Road Assessment, 1999). As in the above example, let the value of output in  $r$  equal 1 million euros and the total transport cost equal 0.1 million euros. Now, however, we assume a monopoly with constant marginal cost, and a higher price elasticity of demand ( $\nu = 3$ ). Again we choose  $p_0 = 1$  such that  $\pi_0 = 0.1$ ,  $q_0 = 1.1$ , and  $D = 1.1^\nu \times 10^6 \times q^{-\nu}$ . The monopoly faces a price elasticity of demand with respect to the manufacturer's cost equal to  $-p/(p + \pi)$ . Note that, while the elasticity with respect to the cost to the customer  $q$  equals  $-\nu$ , the elasticity with respect to the cost to the manufacturer is a bit smaller and decreasing in  $p$  in absolute value. The first-order condition of profit maximization (the so-called Amoroso-Robinson condition) then yields  $p(\nu - 1) - \pi = \nu c$  with marginal cost  $c$ . Inserting  $p_0 = 1$ ,  $\pi_0 = 0.1$ , and  $\nu = 3$  gives  $c = 1.9/3 \approx 0.63$ .

When, after the shock, transport costs are reduced to  $p_1 = 0.05$  we get  $p_1 = 0.975$  by inserting  $c$ ,  $\nu$ , and  $\pi_1$  into the Amoroso-Robinson condition; the cost to the manufacturer falls by 2.5% because the falling transport costs make demand a bit more elastic. The cost to the customer falls by 6.8% ( $q_1 = p_1 + p_0 = 1.025$ ) because both  $p$  and  $\pi$  fall. The total welfare gain now amounts to 139 000 euros (areas A plus B plus D minus area C in Figure 3), which is nearly three times the direct cost saving (area A minus C in the figure). Thus, 83 000 euros go to the consumers, and 56 000 euros to the producer. The reason is, of course, that output

expansion generates an extra welfare (area D) because the manufacturer's cost exceeds the marginal cost.

Three lessons can be learnt from this example:

- The effects of transport cost changes strongly depend on the assumptions about market structure. In fact, for infinitesimal changes the welfare gain just equals the direct cost saving under perfect competition, while the welfare gain can be several times larger (or much smaller and even negative) as the direct cost saving under imperfect competition. Hence, measuring market power, wage rigidity, and other imperfections becomes crucial in a realistic assessment.
- Unfortunately, imperfect markets increase the risk of the arbitrariness of the results. The results may be very sensitive with respect to parameters and behavioral assumptions, which are difficult to check empirically.
- Partial equilibrium results, as produced in the example above, may be seriously misleading. Implicitly, the example is based on the assumptions that other markets are non-distorted, or are characterized at least by a lower relative excess of prices over marginal cost. If all prices exceeded the marginal cost by the same factor in each industry, then the relative prices would not be distorted, and output expansion in the industry under consideration would not generate an extra welfare gain. Hence, general equilibrium analysis is particularly difficult with imperfect markets, but at the same time also especially important.

An elegant way of introducing imperfect competition is the Dixit and Stiglitz (1977) version of monopolistic competition, introduced into spatial equilibrium modeling by Krugman (1991) (see also Fujita et al., 1999). According to this approach, goods and services within each sector consist of a set of diversified symmetrical brands, each produced under increasing returns. Each brand has a negligible weight in total demand, and free entry drives profits to zero. Substitution between brands is modeled by a lower CES (constant elasticity of substitution) nest in production and utility functions.

If combined with the "iceberg"-type transportation technology discussed above, this approach turns out to be comparatively simple in terms of data requirements and computational complexity. This explains its recent popularity, but it comes at a cost: there is no separate parameter controlling the degree of competition in an industry, and the impact of local intensity of competition on monopolistic price mark-ups is excluded by assumption. To what extent project evaluations under this market form deviate from those under perfect competition largely depends on a key parameter, the elasticity of substitution between brands. Bröcker (2001) shows that results using this approach deviate only slightly from what would have been obtained under a perfect competition assumption. This is because the calibrated parameters imply a rather low degree of monopoly.

#### 4.2. Dynamics

Applications of CGE methods to transport issues up to now have been static. One compares snapshots of the economy for a point in time with a given stock of factors and of knowledge. In other fields of application, dynamic models are already state of the art, though their sophistication can differ markedly.

A comparatively simple approach conceives of saving and technical progress as exogenous, and introduces *ad hoc* assumptions regarding migration and investment allocation. An example is the GTAP model (Hertel, 1997), which avoids modeling the behavior of forward-looking agents forming rational expectations. Technically speaking, these approaches require no more than solving a series of static CGEs, one for each period, rather than a single one. From period to period, factor stocks and the state of technology change due to migration and natural demography, capital accumulation, and exogenous technical progress. Such an approach can be used to predict the long-term consequences of transport policies, but the theoretical basis is not very convincing, and a consistent welfare evaluation is impossible.

Much more demanding is to introduce neoclassically derived saving, investment, and migration decisions. A realistic model requires the introduction of adjustment costs for migration and investment. As a consequence, saving, migration, and investment must be obtained from the intertemporal optimization of forward-looking agents. A solution requires the use of the Hamilton or Bellmann formalism, and leads to two-point boundary value problems, which are still difficult to handle for non-linear systems in many dimensions.

Even higher complexity arises with endogenous technical change, the subject of new growth theory (Barro and Sala-i-Martin, 1995). Potentially, this opens an extremely interesting field for future research in transportation economics (Standing Advisory Committee on Trunk Road Assessment, 1999). Technical change results from knowledge production, the main input of which is again knowledge. Inter-regional transfer of knowledge is costly, even with modern telecommunications. To a large extent, knowledge is incorporated in human beings, and knowledge exchange requires face-to-face contact. Hence, inter-regional knowledge flows are influenced to a great extent by the cost of passenger transport. This implies that passenger transport cost may influence the speed and the spatial pattern of innovation and thus have an impact not only on the levels of but also on the rates of the growth of economic activity. There is still very limited knowledge about these issues (for a theoretical treatment see Walz, 1996), with a long way to go until the ideas from recent endogenous growth theory find their way into applied transportation economics.

### 5. An example: the spatial effects of trans-European road networks

The results obtained by a CGE for a multi-regional system in Europe are presented, to demonstrate that the approach is operational and useful in practice. The aim of

the analysis was to evaluate the effects of transport links that are planned or under construction for the Trans-European Transport Network (TEN-T) (European Commission, 2000). The study is confined to the regional welfare effects resulting from the use of the new links for commodity trade. Effects from the construction phase, financing, and maintenance are not considered. Use of the links for other than trade purposes, such as commuting, tourism, and leisure trips are not considered either. The analysis uses a static equilibrium model for a closed economy covering the whole world, subdivided into a large number of regions interacting through commodity trade. Europe (from the Atlantic to the Urals) is subdivided into 800 regions; the rest of world is aggregated into five broad areas.

In each region there is one household, representing all final demand, including public expenditure as well as investment, and two firms, representing the local sector and the tradables sector. The local sector produces for the regional market only, while the tradables sector produces a subset of tradable product varieties for customers in the whole world (including their own region). Customers regard product varieties as imperfect substitutes such that there is a considerable amount of two-way trade (cross-hauling). One can show that the trade flows in the equilibrium solution obey a gravity law, which is well known to fit the data excellently. The market for local goods assumes perfect competition, while the market for tradables is governed by monopolistic competition with free entry according to Dixit and Stiglitz (1977), as explained above. Trade between regions is costly, with the costs depending on transport distances through a given transport network as well as on national trade barriers. The welfare implications of the new transport links are quantified by simulating the effects of transport distance reductions. More details are given in Bröcker (1998b), and an application of the same model to a different issue are given in Bröcker (1998a).

Figure 4 shows a typical result, namely the regional welfare gains from the so-called Helsinki corridors. In 1997 the Third Pan-European Transport Conference in Helsinki defined nine combined rail/road corridors ("Helsinki corridors"), and the Danube inland waterway as the 10th transport corridor (European Commission and TINA Secretariat, 1999). The results in Figure 4 assume that freeways are built along each corridor (except the Danube, of course). Welfare gains are measured by equivalent variations as percentages of regional GDP. They range from losses close to zero up to gains of 3% of GDP per year (darkest shading in the figure). The model is calibrated with 1995 data. The large gain of 3% of GDP is in the Kaliningrad region, the Russian enclave by the Baltic sea, which is suffering from isolation today and would be particularly well connected under the assumed infrastructure scenario. A few urban regions (St Petersburg, Smolensk, and Novgorod) gain between 2% and 3% of GDP. The other gains are less than 2%.

This is just one of many possible applications. Note that one can evaluate not only the impact of infrastructure investment but also the impact of regulations, externality pricing, and other kinds of transport policies. These would all affect

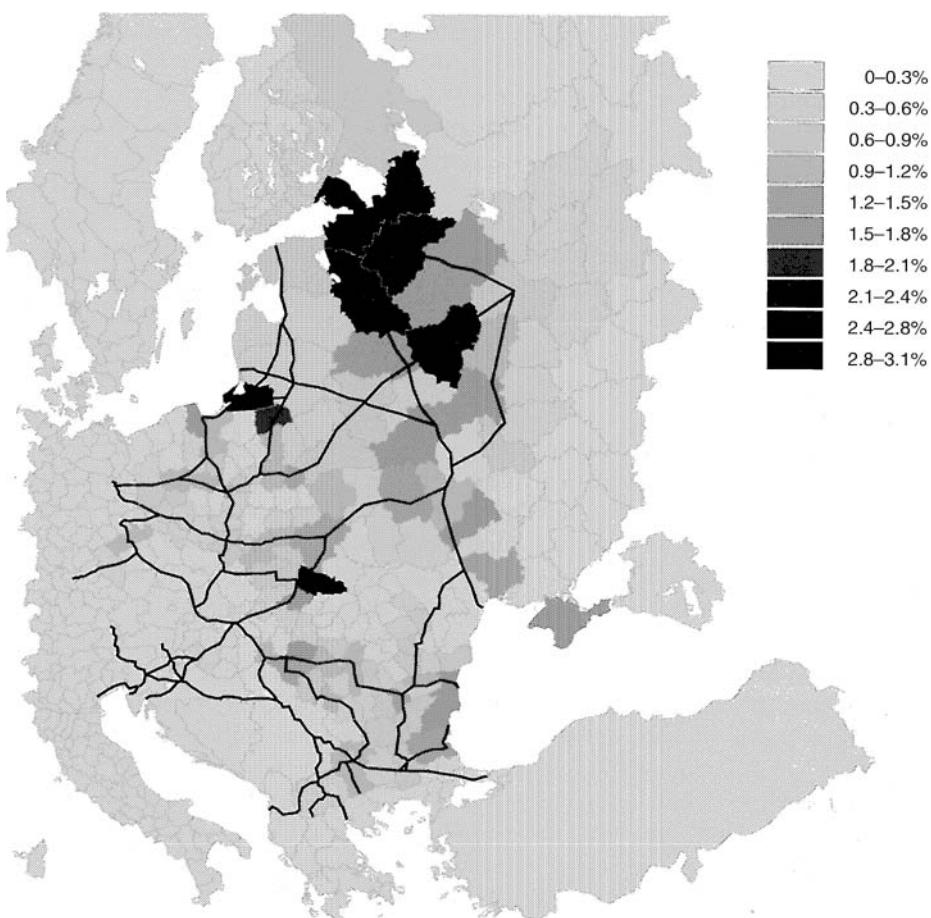


Figure 4. The welfare effects of the Helsinki corridors.

transport costs, goods, and factor prices, and lead to welfare gains or losses varying over space. Introducing a public budget also allows for incorporation of the effects from financing the investments, redistributing revenues, externality pricing, etc.

## 6. Conclusions

Four conclusions can be drawn from the discussion in this chapter on CGE modeling:

- State-of-the-art multi-regional CGE models are a powerful operational tool for analyzing the spatial implications of transport policies. They stand on a firm theoretical basis, and can be implemented with an empirical effort comparable to more traditional methods, taking complex economic repercussions of cost reductions into consideration.
- CGE transport models can be connected well with network equilibrium models, to generate a hybrid system portraying transport and the economic environment.
- Extensions of CGE transport models to more realistic imperfect market economies are well under way. A warning has to be given, however, that the inherent uncertainty can produce varying assumptions on the behavior of firms under imperfect competition.
- The dynamic effects of transport are, in the light of new growth theory, very important, but as yet neither well understood theoretically nor incorporated in CGE models in a satisfying manner.

## References

- Allgower, E.L. and K. Georg (1990) *Numerical continuation methods: an introduction*. Berlin: Springer-Verlag.
- Ando, A. and T. Shibata (1997) "A multi-regional model for China based on price and quantity equilibrium," in: M. Chatterji, ed., *Regional science perspectives for the future*. London: MacMillan.
- Armington, P.S. (1969) "A theory of demand for products distinguished by place of production," *International Monetary Staff Papers*, 16:159–176.
- Barro, R.J. and X. Sala-i-Martin (1995) *Economic growth*. New York: McGraw-Hill.
- Bröcker, J. (1998a) "How would an EU-membership of the Visegrád countries affect Europe's economic geography?" *Annals of Regional Science*, 32:91–114 (erratum (2000), 34:469–471).
- Bröcker, J. (1998b) *Spatial effects of new transport links: preliminary results from a spatial computable general equilibrium analysis. Diskussionsbeiträge aus dem Institut für Wirtschaft und Verkehr, Technische Universität Dresden*, No. 4/98. Dresden: Technical University.
- Bröcker, J. (2001) "Spatial effects of transport infrastructure: the role of market structure," in: J.R. Roy and W. Schulz, eds, *Theory of regional competition*. Baden-Baden: Nomos.
- Bröcker, J. (2002) "Passenger flows in CGE models for transport project evaluation," in: *ERSA Congress 2002*, Paper. Dortmund.
- Bröcker, J., A. Kancs, C. Schürmann and M. Wegener (2001) *IASON deliverable 2: methodology for the assessment of spatial economic impacts of transport projects and policies*. Delft: TNO Inro (<http://www.inro.tno.nl/iason/Documents/Project/iason/Deliverables/IASON%20D2.pdf>).
- Buckley, P.H. (1992) "A transport-oriented interregional computable general equilibrium model of the United States," *Annals of Regional Science*, 26:331–348.
- Conrad, K. (1997) "Traffic, transportation, infrastructure and externalities," *Annals of Regional Science*, 31:369–389.
- Conrad, K. and St. Heng (2000) *Financing road infrastructure by savings in congestion costs: A CGE analysis*. Universität Mannheim, Institut für Volkswirtschaftslehre und Statistik, Beiträge zur angewandten Wirtschaftsforschung, No. 579-00. Mannheim: University of Mannheim.
- Diewert, W.E. and T.J. Wales (1987) "Flexible functional forms and global curvature conditions," *Econometrica*, 55:43–68.
- Dixit, A.K. and J.E. Stiglitz (1977) "Monopolistic competition and optimum product diversity," *American Economic Review*, 67:297–308.

- European Commission (2000) *The common transport policy, sustainable mobility: perspectives for the future. Commission Communication to the Council, European Parliament, Economic and Social Committee and Committee of the Regions.* Brussels: EU (<http://europa.eu.int/comm/transport/themes/mobility/english/en1.pdf>).
- European Commission and TINA Secretariat (1999) *TINA draft final report.* Brussels: EU.
- Friesz, T.L., Z. Suo and L. Westin (1998) "Integration of freight network and computable general equilibrium models," in: L. Lundquist, ed., *Network infrastructure and the urban environment: advances in spatial systems modelling.* Berlin: Springer-Verlag.
- Fujita, M., P. Krugman and A.J. Venables (1999) *The spatial economy: cities, regions, and international trade.* Cambridge: MIT Press.
- Gazel, R. (1996) "Free trade agreements and interregional labor migration: the case of the US and Canada," *Annals of Regional Science*, 30:373–390.
- Haddad, E. and G.J.D. Hewings (2000a) "The short-run regional effects of new investments and technological upgrade in the Brazilian automobile industry: an interregional CGE analysis." Unpublished.
- Haddad, E. and G.J.D. Hewings (2000b) "Transportation costs and regional development: an interregional CGE analysis." Unpublished.
- Harberger, A.C. (1962) "The incidence of the corporation income tax," *Journal of Political Economy*, 70:215–240.
- Hertel, T.W. (1997) *Global trade analysis: modelling and applications.* New York: Cambridge University Press.
- Hirte, G. and J. Genosko (1988–1989) "Regionalisierte empirische allgemeine Gleichgewichtsanalyse: eine Einführung mit einem einfachen Modell für die Bundesrepublik Deutschland," *Jahrbuch für Regionalwissenschaft*, 9/10:32–56.
- Horridge, M. (1994) "A computable general equilibrium model of urban transport demands," *Journal of Policy Modeling*, 16:427–457.
- Hussain, I. (1996) *Benefits of transport infrastructure investment*, Umeå Economic Studies No. 409. Umeå: Umeå University.
- Hussain, I. and L. Westin (1997) *Network benefits from transport investments under increasing returns to scale*, Umeå Economic Studies No. 432. Umeå: Umeå University.
- Johansen, L. (1960) *A multi-sectoral study of economic growth.* Amsterdam: North-Holland.
- Jones, R. and J. Whalley (1988) "Regional effects of taxes in Canada: an applied general equilibrium approach," *Journal of Public Economics*, 37:1–28.
- Knaap, T. and J. Oosterhaven (2001) *The welfare effects of a new infrastructure: an economic geography approach to evaluating a new Dutch railway link.* Groningen: University of Groningen.
- Krugman, P. (1991) *Geography and trade.* London: MIT Press and Leuven University Press.
- Madden, J.R. (1992) *The theoretical structure of the federal model*, CREA Paper No. TS-02 (revised). Hobart: Centre for Regional Economic Analysis, University of Tasmania.
- Mayeres, I. (2001) *Equity and transport policy reform*, Working Paper Series No. 2001-14. Leuven: Faculty of Applied Economic Sciences, Energy, Transport and Environment, Catholic University.
- Samuelson, P.A. (1954) "The transfer problem and transport cost, ii: analysis of effects of trade impediments," *Economic Journal*, 64:264–289.
- Sheffii, Y. (1985) *Urban transportation networks.* Englewood Cliffs: Prentice-Hall.
- Shoven, J.B. and J. Whalley (1984) "Applied general-equilibrium models of taxation and international trade: an introduction and survey," *Journal of Economic Literature*, 22:1007–1051.
- Standing Advisory Committee on Trunk Road Assessment (1999) *Transport and the economy.* London: UK Department of the Environment, Transport and the Regions.
- Trela, I. and J. Whalley (1986) *Regional aspects of confederation. Royal Commission on the Economic Union and Development Prospects for Canada*, Vol. 68. Toronto: University of Toronto Press.
- Ueda, T., T. Tawai and Y. Hayashiyama (1999) "Ex-post evaluation of Japanese high-speed transport systems," in: *International Symposium on Structural Changes in Transportation and Communications in Knowledge Society: Implications for Theory, Modeling and Data*, Paper. Boston: Boston University.
- Venables, A.J. and M. Gasiorek (1998) *The welfare implications of transport improvements in the presence of market failures*, SACTRA Working Paper A, Revised 4/1/98. London: UK Department of the Environment, Transport and the Regions.

- von Thünen, J.H. (1826) *Der isolirte Staat in Beziehung auf Landwirtschaft und Nationalökonomie*. Hamburg: Friedrich Perthes.
- Walz, U. (1996) "Transport costs, intermediate goods, and localized growth," *Regional Science and Urban Economics*, 26:671–695.

***Part 4***

**DATA**

## **SPATIAL DATA ISSUES: A HISTORICAL PERSPECTIVE**

**PETER R. STOPHER**

*The University of Sydney*

### **1. Introduction**

Travel is a phenomenon that takes place in both time and space, as was clearly expounded by Hagerstrand (1970) and Lenntorp (1976). This work is also discussed at a little more length in Chapter 36 of this book. The temporal aspects of travel are relatively easily handled, and only brief mention of them is made in this chapter, but the spatial aspect of travel presents a number of challenges to the transport analyst. Transport planners need to know where travel will take place, and the amount of travel at different locations. In developing models to define this, transport planners need to know where people live and the travel that they undertake, by geographic location, and where the destinations are to which people travel. Traffic engineers need to know where congestion occurs in the system, where traffic signals and other traffic control devices are, and where there are changes in the supply of capacity for traffic. Bus planners need to know where bus stops are located, and where people who use the services originate, arrive, and where they transfer between services. Rail planners need to know at which stations passengers board and alight, where transfers are made, and how people get to and from the rail station. In short, all aspects of transport planning, traffic engineering, public transport operation and management, investment planning, etc., require information about various spatial aspects of the transport system.

While geographic information systems (GIS), as discussed in subsequent chapters of this book, offer enormous capabilities to understand and analyze spatial aspects of transport, they have been implemented only to a limited degree thus far, and spatial issues continue to confront the transport planner. The principle issue, thus far, in spatial representation and analysis of travel and traffic must be that of spatial aggregation. In the balance of this chapter, we present issues relating to spatial aggregation of residences and employment, then spatial aggregation of the transport network, and interactions between the two. We then review the potential directions for resolving some of these issues.

The discussion in this chapter focuses principally on the urban context, because this has currently received by far the greatest attention within the profession. However, almost everything stated here also applies equally well to state/province, national, and international studies of travel. It is principally the scale of the geography that varies.

## **2. Traffic analysis zones**

Since the inception of urban transportation planning in the early 1950s, the study region has been divided into traffic analysis zones (TAZs). Originally, this was a necessity because of the limited capability of computers to handle large matrices, and the need to find a way to represent the entire region under study in the computer for analysis purposes. In the beginning, the number of zones was limited to the maximum matrix size that computers dating from the 1950s could handle. The result was the use of a small number of rather large zones, typically of the order of 150–250 zones for an entire metropolitan region. This number of zones gave rise to a matrix containing from 22 500 to 62 500 cells (US Department of Commerce, 1964; Stopher and Meyburg, 1975). These zones were generally defined to have reasonably consistent numbers of trips originating and arriving in them. Therefore, zones in the central business district (CBD) were geographically quite small, while those in outlying areas were very large. This is shown in Figure 1. As Stopher and Meyburg point out, there were often at that time two zoning systems – one was used for data collection, and consisted of some amalgamation of units that related in some way to a sampling frame or other procedure used in the data collection, an example of which is shown in Figure 2, while the second was the analysis zones that would be defined on the basis of homogeneity and transport criteria, outlined below, of which there is an illustration in Figure 3 for the same region as Figure 2.

At least two features were recognized early in the process: first, that zones needed to be as homogeneous as possible, because they would be represented in the subsequent modeling principally by their means, and second that zones should act as catchment areas for trip ends, and therefore should be bounded by major roadways, or barriers to movement, and should not have major transport thoroughfares or barriers intersecting the zone. As experience with travel demand modeling began to be accumulated, other requirements became apparent for TAZs. In addition to the need to bound TAZs by major transport corridors or barriers to movement, to prevent major thoroughfares or barriers from cutting across the TAZ, and to try to achieve as much homogeneity as possible of the households and businesses in a zone, it also became apparent that there needed to be some relationship between zones and the spatial units used by the census, because this was the source of much of the data required about households. In the

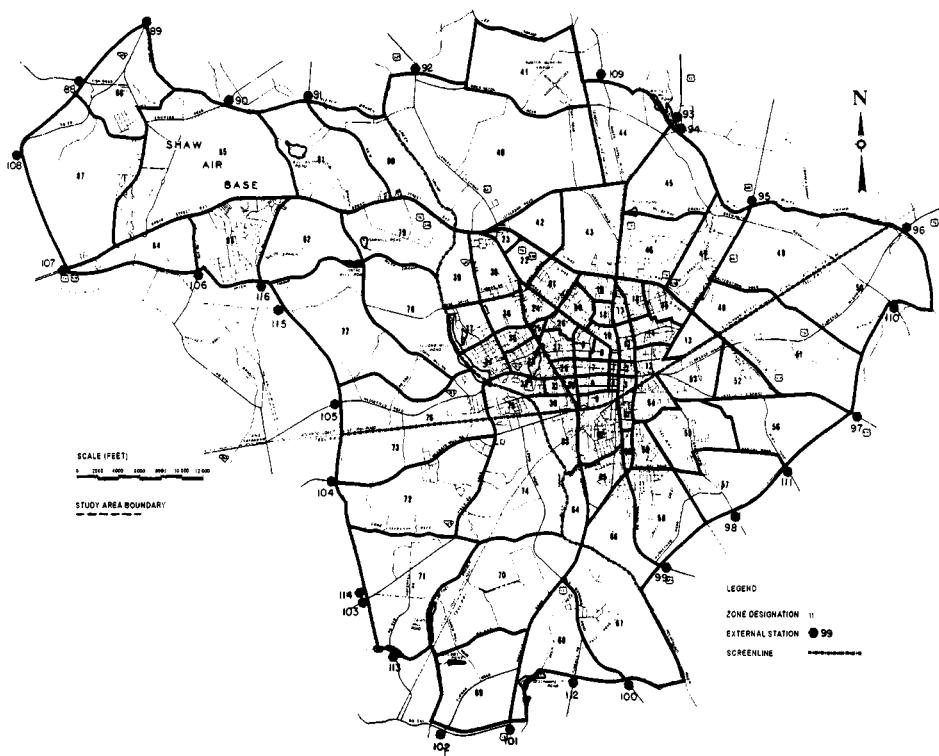


Figure 1. Typical recommended zoning system for a regional study.

USA, this meant that census tracts, which through at least the first 30 years of urban transport planning represented the smallest unit of census geography for which data would normally be released and available for transport planning, must map into the TAZs. Yet a further requirement was that zone boundaries should map into major and minor civil division boundaries, so that complete groups of zones would represent each minor civil division within the study area, while the region as a whole might be bounded by the boundaries of major civil divisions.

Early travel demand modeling was characterized by models that expressed zonal totals as their dependent variables. Thus, the early trip generation models estimated the total numbers of trips produced by and attracted to each zone, early modal split models estimated the numbers of trips produced by and attracted to zones by each of car and public transport, the trip distribution models estimated the total numbers of trips traveling between pairs of zones, and traffic assignment models estimated the volume of trips on a link of the network. Two results arose from this type of modeling. First, the necessity for means to represent the values in

the zones became extremely important. This, in turn, required that the distributions of population and employment characteristics obey the central limit theorem, so that the mean is a reasonable representation of the distribution. Second, the models are tied to the zone system at the time that the models are estimated, because the relationships are dependent on the precise ways in which the zones have been defined. Adding new zones will result in a requirement for extrapolations outside the range of the model estimation. Changing the boundaries of existing zones will invalidate the model relationships, because they are defined by the zones at the time of the model estimation.



Figure 2. Sampling districts for the London Traffic Survey, 1963.



Figure 3. Traffic analysis zones for London in 1963.

McCarthy (1969) and Kassof and Deutschman (1969) investigated these requirements and concluded that there were substantial and significant problems in the use of zones and zone total models. Among McCarthy's findings were that most zones in use up to that time were highly heterogeneous, with much higher variances within zones than between them, and that the distributions of most variables in zones were skewed, so that the mean was not necessarily a good representative value for the zone in model estimation. Additional findings on the use of zonal totals included the problem that variability in zone sizes also affected the accuracy of the resulting models. At that time (and still to the present), it was

customary to use small zones in the CBD, and to allow the zones to become progressively larger as one moved further out from the center of the region. While this would seem to be appropriate to maintain a consistent population size in a zone, it usually did not do so, with CBD zones being much smaller geographically and larger in employment, particularly, and zones at the periphery of the region being geographically very large, but having small population and employment totals. This, it was found, gave rise to problems of accuracy.

Subsequently, there was increasing development of models that were estimated on more disaggregate information, such as trip generation rates models, which estimated household-based trip rates. This resulted in less concern with some of the aspects of the zone system, but did not relax the requirements for homogeneity, nor for zones to represent catchment areas for trip ends. While these models often did not use any zonal aggregation for the estimation of model parameters, they were still used to produce values for zones, because the end result of the modeling procedure was a requirement for assigned volumes on street segments. Achieving these assigned volumes still required that zonal numbers of trips by purpose, and mode were required – forming a set of trip tables that would be assigned to a network, using zone centroids as the origin and destination points for all trips.

Since the early days of urban transport planning the capabilities of computers have increased, and the matrix limitations of the early 1950s computers no longer exist. In the 1970s, most transport-planning studies used as many as 750–1000 traffic analysis zones, and the distinction between sampling zones and analysis zones was gradually lost. The increasing numbers of zones should have led to improvements in zone homogeneity, but generally did not. The reason for this is that the requirement to map into census geography became, if anything, more significant, so that TAZs now began to be defined as either single census tracts or as amalgamations of two or three tracts. Outside the USA, the restrictions of census geography were highly varied, so that some countries found similar problems to the USA while others did not. The primary problem with following census geography is that it does not usually follow the rules desired for TAZs. Although it does not happen often, census geography may cross major thoroughfares, rather than using them as boundaries. Homogeneity is not a major concern of census geography. As the numbers of zones increased in the 1970s and 1980s, the requirements to match into the smallest available census geography continued to be paramount, particularly because methods to split the information from the census across census geography boundaries did not exist.

Today, the number of zones continues to grow, with some regions using as many as 5000 zones, although most use around 1000–2000. New problems arise with the increase in numbers of zones. While smaller zones will, *ceteris paribus*, lead to more homogeneity of the zones, the number of cells in the matrices becomes very large (a zone system with 5000 zones will produce a trip matrix with 25 million

cells), and problems of sparse matrices become the major issue. For example, consider an urban region for which a household travel survey has been undertaken of 3500 households, and where the number of TAZs defined is 5000. Assume that the population of the region is 4 million people in 1.75 million households. The average zone will have 350 households in it. Assume that the average number of trips made by a household in 1 day is 10, then the survey will produce about 35 000 trips that will be spread over the 25 million cells of the trip matrix. This means that, if the survey measured one trip between any origin and any destination, only 0.14% of the cells of the matrix would receive even one trip. If the survey produces multiple trips for some origin–destination pairs, then perhaps as little as 1/10 of 1% of the cells of the matrix will have trips present. Even if it were possible to take a census of the regional population, there would be no more than about 17.5 million trips from 1 day. This will still create a sparse matrix. Again if one assumed that travel was so diverse that no two persons made a trip between the same zone as the origin and the same one as the destination, then there would be 30% of the cells in the matrix that would be empty. If most trips from an origin to a destination produced another trip returning to the first origin, then the number of cells that are empty would have to double. This means that the identification of which cells in a matrix are genuinely empty, as opposed to those that are empty because of the sparseness of survey information, becomes a critical issue. This has yet to be resolved.

The duality of zoning systems has, however, not disappeared. Instead of having separate zoning systems for sampling and for analysis, a new zone system is introduced when there is a land use model in the model set. Land use models are generally applied at much higher levels of aggregation than travel-forecasting models. In situations where there may be 1000 traffic analysis zones, there may be only 150 or so land use zones. In many cases, the land use zones are constructed as aggregates of the traffic analysis zones, but this is not always the case. In any event, because the land use information is required to run the travel forecasting models, the land use model outputs to the land use zones have to be disaggregated to the smaller TAZs. Usually, this disaggregation assumes either that distributions are uniform among the TAZs that make up the land use zone, or that population and employment will be distributed proportionately to the current distributions. In either case, this adds further complexity to the spatial analyses. In addition, it raises an issue of accuracy. If land use forecasts are made only at a “super-zone” level, with relatively poor information on how to allocate the resulting population and employment forecasts to the detailed TAZs, then the potential accuracy of the travel-demand models is compromised.

GIS provide the means to split data among multiple TAZs, when several TAZs fall into a single unit of census geography. However, most splitting procedures in GIS assume that the underlying distribution of each characteristic is uniform. In terms of the types of characteristics of interest in transport planning, this is

unlikely to be a good assumption. Rather, it is likely that the characteristics will be distributed unevenly, and probably in a skewed distribution. At this point, one must start to consider whether there is any virtue in a spatial aggregation of households. Possibly, it is now better to use individual households in the modeling and to use random sample enumeration as the procedure to aggregate to total regional trips.

### **3. Traffic networks**

The second spatial component of the system is the transport network. In GIS terms, the TAZs discussed in the previous section represent a polygon layer in a GIS, while the network represents a line layer. In the early applications of urban transport-planning models, the network was represented in the computer as a system of points (nodes) and lines (links). The points or nodes represented the intersections of multiple lines in the network, while the lines or links represented the traveled way between intersections. An illustration of this is shown in Figure 4. Computer storage and processing limitations also limited the detail of the network that could be used in the early years of transport planning. Generally, the number of nodes had to be kept to around 1000, with no more than about 2000 links. These limits relaxed progressively as computing power increased. However, the guidance developed in the 1960s has generally been held to in subsequent years. Under this guidance, the network consists of freeways, expressways, arterial roads, and collectors and distributors. Local streets are not included in most urban area networks.

It is also customary to represent most streets by a single two-way link. Apart from the obvious exception of one-way roads, the exception to this is usually that freeways and other divided roadways are coded as pairs of one-way links, particularly if there are grade-separated interchanges, with ramps that lead traffic on to one or other direction of movement. Nevertheless, there are many instances where, even today, freeways are to be found in networks represented as a single link, with no detail on the ramps, and freedom to interchange at a single node that represents the intersection of the freeway with any surface street. This presents serious issues in realistic network coding, particularly in the event that an interchange is incomplete and does not allow all movements on and off the freeway. The presence of such coding is probably a throwback to early days when there simply was not the computer storage to handle these situations, and where limited access roadways were not common. Otherwise, there is no reason from a computer capacity viewpoint to make such erroneous simplifications in network representation.

As with the zone systems, the level of aggregation of the network presents some problems. All traffic is assumed to travel on the coded network. As a result,

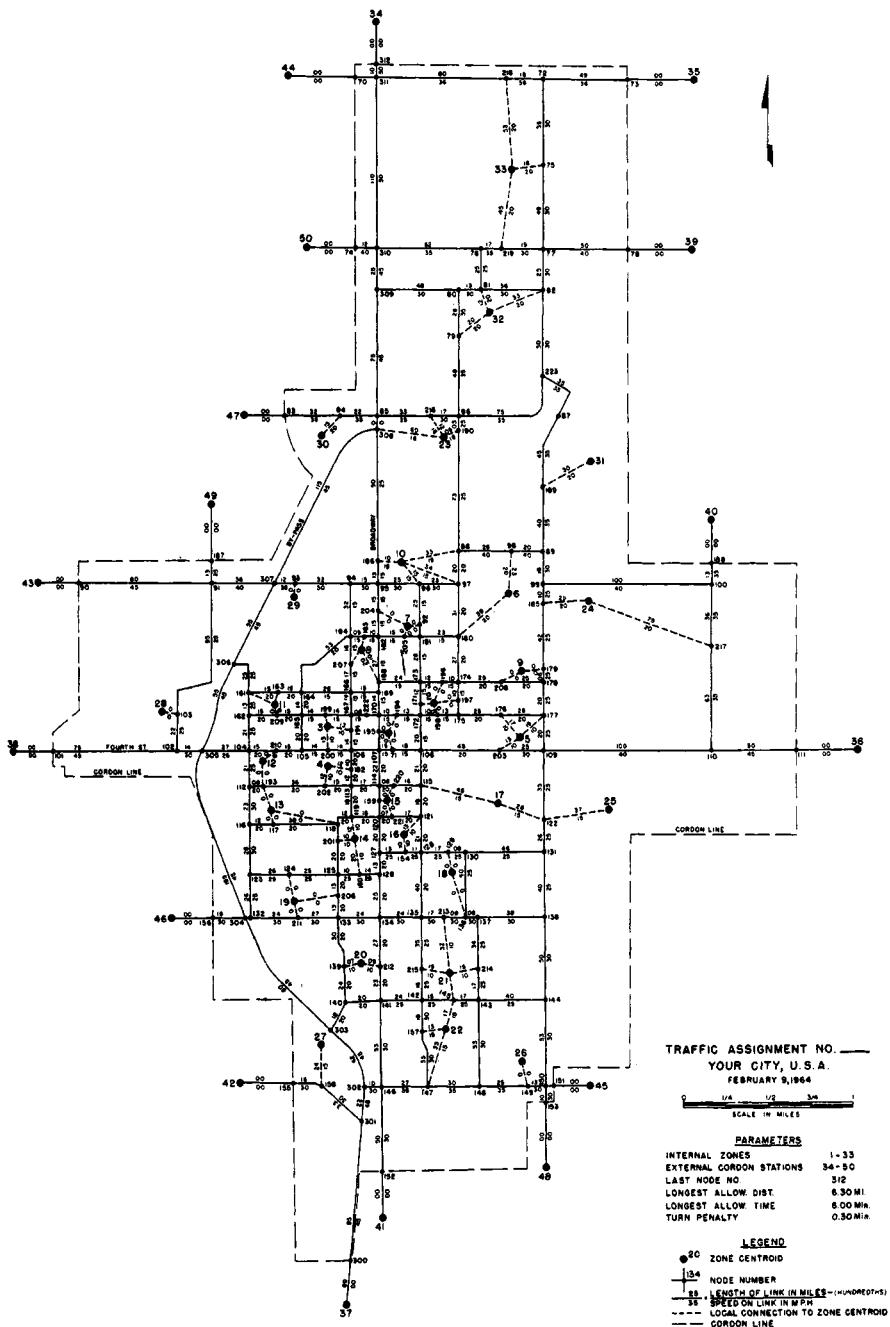


Figure 4. Typical planning street network from the 1960s

correct identification of what constitutes the arterials, collectors, and distributors is paramount. If traffic is actually using what are considered to be local roads for through movements to any significant extent, then the volumes of traffic shown on the coded network will be too high, while problem areas in the network may not be identified, because they may be occurring on roads that are not present in the coded network. Another problem with networks in dedicated transport-planning software is that links are straight lines connecting intersections, and correct geography of the street system is usually absent. This means that the physical length of links in the network is often not correct, and geometry of the street system is not represented.

Another issue in network coding is that nodes are simply anchor points for the ends of links, and do not possess attributes in standard network terminology. The links possess attributes, such as length, number of lanes, presence or absence of parking, etc. Most importantly, links possess the attribute of capacity, which is somewhat of a problem, because it is actually the intersections (nodes) whose capacity is usually the limiting capacity for a link. Thus, there is a transfer of the intersection capacity to the link in standard network coding procedures.

### *3.1. Bus networks*

Bus networks were initially created completely separately from the highway network. Of course, in the earliest years of transport planning, only highway networks were created, because the approach to transport planning was really a highway-planning approach. However, during the 1960s, planning for public transport began to arise as an issue, and bus networks needed to be developed. Computer capability was still limited, and initial networks were independent from the highway network. To the extent that congestion from the highway network was needed to be reflected in the bus network, to slow bus speeds, it had to be transferred manually from the highway network. In the 1970s, the capability was developed to create a bus network on the street network developed for the highway side of the planning, and to automatically reflect the levels of highway use in bus speeds. Bus and rail networks (and ferry, where appropriate) are now most usually created as networks that are based on the highway network, with the addition of links and nodes that represent the traveled ways for dedicated public transport, such as rail lines, busways, and other similar facilities.

One of the issues that affect transit networks is that there is not an established relationship between capacity and travel time, as there is on the road network. Therefore, there is not usually a capacity attribute in a transit network. This also means that there is not a capability to analyze the network to reflect the effects of demand exceeding the supply of public transport capacity. Second, transit networks create several problems for current network representation, such as

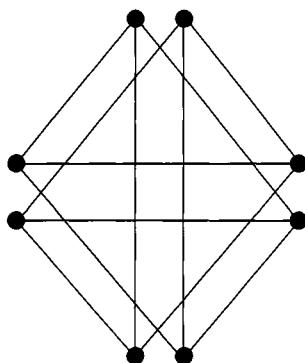


Figure 5. Micro-network of an intersection.

there being different operating speeds along the same link for a local bus, a limited stop bus, and an express bus. Also, when bus routes involve short turns, loops, and other similar features, there are a variety of problems that arise for representing the routes realistically in a computer network. These are issues that tend not to be handled well in most software applications, but are too complex to attempt to handle in the context of this chapter.

### 3.2. Micro-networks

Other types of networks can also be constructed. One of these is a micro-network, in which a single intersection may be provided in great detail. In micro-networks, the standard approach of nodes and links is reversed, so that the intersection is represented by a number of links, and the traveled route between intersections is represented by a node. In these networks, the links represent each of the possible movements through the intersection. A simple four-way intersection may thus be represented by 12 links connecting the eight nodes that represent the approach streets, by direction, as shown in Figure 5. The next intersection would then start from the nodes at the extremity of this figure. Again, the links have attributes, only this time these attributes relate to movement through the intersection, and the nodes represent the location of the traveled way between the intersections.

## 4. Interactions between zones and networks

Zones and networks need to be related to one another, because the zones are the sources and sinks for trips, and the trips travel over the network to get from one

zone to another. Traditionally, this has been done by representing the zone by a point location for purposes of interrelating it with the network. This point location is usually referred to as a zone centroid, and is thought of much as the physical centroid or center of gravity of the zone. In turn, the zone centroid is connected to the network by links that are called centroid connectors. These links, unlike the rest of the network, do not represent actual street links on the ground. Instead, they represent an aggregation of the local roads and streets that provide access to the collectors, distributors, and arterial streets.

In reality, the centroid and its connectors must represent the travel that comes from or ends at a variety of locations within the zone, using available streets within the zone. As a result, it is customary to make each centroid connector within a zone the same length (in terms of time and distance), regardless of the physical positioning of the centroid. Clearly, the larger the zone, the more inaccurate will be the use of a centroid as the point source and sink for trips, and the more inaccurately will the centroid connectors represent the travel times related to getting into and out of the zone. When zones are physically very small, the centroid becomes a more accurate representation of where trips end, and the connectors are likely to represent more accurately the underlying local streets.

In a public transport network, the centroid and its connectors raise some additional problems. When public transport is available only on one side of a zone, parts of the zone may be within walking distance of stops, while other parts are too distant to be considered as walking access. Nevertheless, the centroid connectors will probably show the centroid to be within walking distance. This is usually less problematic with bus networks, where service is less often available only to one side of a zone, but is a serious potential problem for rail networks, where one station may be all that is available on one edge or corner of a zone. In such a case, there is no accurate way to portray the range of access times and options that would be used from the zone. A strategy that is often used is to define for each such zone a percentage of the population within walking distance, and use average walking times out of the zone for this group, and to assume that the remainder of the zone population would have to access the train using either bus or car. Again, this problem can be solved by making zones smaller. The smaller the zone, the less likely it is that part of the population will be too distant from the public transport stop or station to walk.

#### *4.1. Zone size and networks*

One of the issues that has limited zone size in the past has been that of network detail. Given the desire to bound zones by streets that are in the network, zones cannot be smaller than the level of detail of the network. At the same time, the

solution to many of the problems confronting planners with spatial data is that zones tend to be too large. The potential to use GIS as the platform for transport planning offers the possibility of reducing zone sizes significantly. This becomes possible, if the underlying street map is defined as the network, along with all of its attributes. In this case, zone size is limited only to the smallest block of buildings or land contained by local streets. This is probably too fine a level of detail for most planning purposes. However, by using local streets in the network and defining zones that are small numbers of blocks, many of the problems of large zone size, imprecise centroid locations, and unequal centroid connectors should be resolved.

Another option, as mentioned previously, is to work toward abandoning the zonal basis of transport planning and move to using individual households. In this case, each household and each business needs to be connected to the network, and a fully disaggregate set of models would be used. This would require abandoning such models as the gravity model of trip distribution, which is an aggregate model, based on trip length frequency distributions and origin–destination matrices, and replacing it with a destination choice model that operates at the level of the individual household or person. GIS have the capability to provide paths from any point in the geographic space of the region to any other point, provided only that these points are connected to the street network. What may be required to do this is to develop programs that would create connections from every land parcel or group of parcels lying inside an area bounded by local streets.

#### *4.2. The use of a GIS as a network platform*

The use of a GIS in which to build the networks holds promise to remove several of the problems discussed here. First, a GIS is not limited in the size of the network that can be developed. Limitations on the numbers of nodes and links in a network are no longer an issue in a GIS. There is no reason why the entire street system may not be used for the network. However, this does raise some new problems. There is now increasing availability of properly projected street maps for GIS applications in most urban areas around the world, particularly in more developed countries. However, street maps generally do not contain the information that is essential for the transport planner to have for a planning network. Attributes, such as capacity, the number of lanes, the presence of parking, facility type, area type, and posted speed limits are generally not included in GIS maps. Adding such data for the entire network can be a major data collection effort. Even when such data may be known, adding it to the street segments is not a trivial procedure. If this is done, however, then the street map becomes a prime candidate to provide a detailed and realistic network for modeling purposes. Clearly, this network will be much richer than any networks

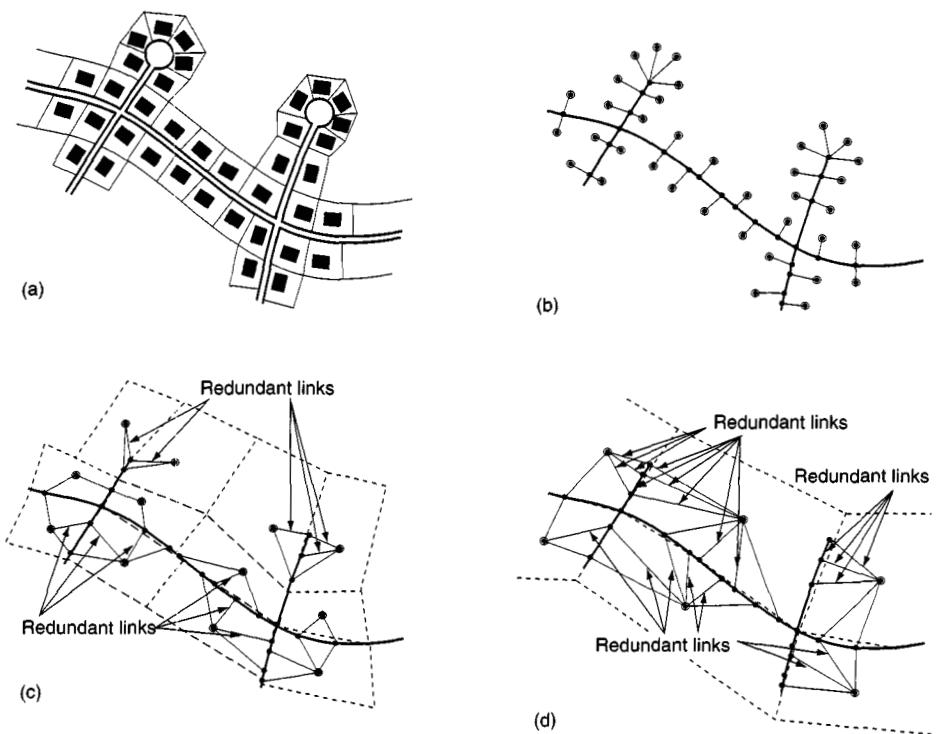


Figure 6. Representation of centroid connectors at different levels of aggregation.

that have been used in the past. However, a question that needs to be addressed here is whether such richness is actually useful.

#### 4.3. Network detail and zone size

There is an important relationship between network detail and zone size. As already noted, each zone is represented in the network context as a single point – the centroid. No matter what level of detail is used in the network, there will be a need to connect each zone centroid to the network through centroid connectors. In aggregate regional representations, such as are common in urban transport planning, these connectors represent the underlying network of local roads that provide access to individual land parcels. If one considered a highly disaggregate representation of the region, in which each parcel of land was a zone (Figure 6a), and every road link on the ground was represented in the street network, the centroid connectors would become the driveways from the land parcels to the

roadway along which the parcel is located, as shown in Figure 6b. At this level, one could justify including every street in the network, because each parcel would produce a separate zone centroid with a single centroid connector (or two if there is a second access location from the parcel). The reason that this can be justified is that there is the potential for at least one trip to need to travel over each link of such a detailed network.

However, if the parcels begin to be aggregated, even if only into groups of two to a few parcels, it would no longer be justified to retain a centroid connector for every driveway, as shown in Figure 6c. This would, in fact, create a situation of redundancy in centroid connectors. Redundancy arises because only those centroid connectors that connect to extreme points in the network, relative to the grouping of parcels, will ever be used in creating a path into and out of the network. Redundant links are shown in Figure 6c. Consider also a residential area containing one or more cul-de-sacs. Centroid connectors that connect to various points along the dead-end section of a street are also redundant, as shown in Figure 6d, in that only a connector to the nearest point on the cul-de-sac link to where it is connected to the rest of the street network will actually be traveled on, if any type of minimum time or cost path is constructed in the network. This means that not only are additional centroid connectors redundant, but also the dead-end link will be redundant from the node to which centroid connectors are inserted to the end of the dead end. This link would never be used by any trip in the network.

From considering these issues, it should begin to become clear that there may be considerable redundancy in a network that attempts to describe all local roads, particularly if there is some spatial aggregation of the land parcels that are served by the network. There is, in fact, virtually no way that links of the road system that fall within an analysis zone will be used, unless there are centroid connectors connected to them, which would generally seem to be inappropriate. It must be concluded, therefore, that the richness of network detail that is offered by a full GIS of the street system would be unhelpful if there is any degree of spatial aggregation of the land parcels served by the network. A means by which to pare the network efficiently to that portion that is required for good analysis and modeling has not yet been devised. It is probably a crucial step to making the full potential of GIS available to transport-planning applications.

## 5. Conclusions

Transport analysis requires spatial information. Methods developed over 50 years ago to handle the spatial context of travel, particularly but not limited to urban areas, resulted in the creation of spatial aggregations of residences and employment locations into zones and other more aggregate units. In addition, the network of streets was represented by links and nodes, representing the segments

of streets and the intersections of the streets, respectively, and with the network aggregated to freeways, arterial streets, and collector/distributor roads. While there is substantial movement toward basing transport planning on a GIS platform, the use of spatial aggregations of the planning region continues, as, to a large extent, does the use of an aggregated street network.

A number of issues arise in the spatial aggregation that is common in transport planning, especially that there are conflicting needs between supplementary data from such sources as the census and the ideals of what would define a traffic analysis zone for the purposes of good travel-demand modeling. In addition, it has been noted that more than one zone system may be required – one for land use modeling and one for travel-demand modeling, but where the two systems need to be able to be mapped into one another. While it is also clear that lower levels of aggregation will lead to less loss of information, and therefore a potential to develop more accurate forecasting models, this chapter has pointed out the problems that arise in disaggregating to larger and larger zone systems, particularly relating to the potential sparseness of trip matrices and the problems of achieving accurate input data on population and employment.

This chapter has also emphasized the links between network level of detail and the zone system. While GIS may offer the capability to include every street segment that exists on the ground, the level of detail produced is well beyond what could be useful for almost any level of aggregation of households, therefore raising the issue of how to reduce the detail of the network in an efficient manner.

Transport is a spatial phenomenon and must be analyzed with due regard to the spatial context. Issues of how to represent the spatial nature of transport are not trivial, and much yet remains to be done to achieve the best balance of detail and aggregation.

## References

- Hagerstrand, T. (1970) "What about people in regional science?" *Papers of the Regional Science Association*, 24:7-24.
- Kassof, H. and H.D. Deutschman (1969) "Trip generation: a critical appraisal," *Highway Research Record*, 297:15-30.
- Lenntorp, B. (1976) *Paths in time-space environments: a time geographic study of movement possibilities of individuals. Lund studies in geography*. Lund: CWK Gleerup.
- McCarthy, G.M. (1969) "Multiple regression analysis of household trip generation – a critique," *Highway Research Record*, 297:31-43.
- Stopher, P.R. and A.H. Meyburg (1975) *Urban transportation modeling and planning*. Lexington: Heath.
- US Department of Commerce (1964) *Traffic assignment manual*. Washington, DC: US Government Printing Office.

*Chapter 18*

## LINKING SPATIAL AND TRANSPORTATION DATA

BRUCE D. SPEAR

*US Federal Highway Administration, Washington, DC*

### 1. Introduction

Transportation is inherently a geo-spatial activity, involving the movement of people and/or things from one geographical location to another. Not surprisingly, therefore, much of the data needed to support transportation analysis, planning, and operations are associated with spatial location. Spatial data used in transportation include: characteristics of the origins and destination locations for trips; descriptions and measures of connectivity and impedance for alternative paths between origins and destinations; locations and descriptions of features and points of interest along a transportation network, and the dynamic location of transport vehicles moving between origins and destinations.

Given the strong linkage between transportation and geo-spatial data, it seems that the transportation community would be an early adopter of geographic information systems (GISs) and other geo-spatial technology. This has not been the case. Transportation planners, in particular, have been relatively slow to embrace GISs as an integral component of their analysis toolkit. Even today, a GIS is often used merely as a medium for presenting transportation data or the results of transportation analyses on a map, while its real strengths for data integration and geo-spatial analyses go underutilized.

The reasons for this are partly historical and partly due to differences in how geo-spatial data are treated in GIS software versus transportation network models. This chapter explores both the historical context and differences between geo-spatial features used in GISs and network models. It concludes with a description of specialized transportation data structures that must be incorporated into GISs to fully address transportation application needs.

### 2. GISs and transportation models – a US historical perspective

#### 2.1. *Origins of GISs*

GISs and transportation models experienced parallel paths of development, tied closely to the development of computer technology. The history of GISs dates

back to the 1960s, as typified by work conducted at the Harvard Graduate School of Design's Laboratory for Computer Graphics and Spatial Analysis (Burrough, 1990). Early applications of GISs focused on area-based analyses, principally the ability to overlay multiple layers of attributes associated with areas (e.g. population density, demographic characteristics, and land use) in order to create composite, multivariate measures that would better discriminate between parts of a larger region. Early computer technology was slow and expensive to operate (by today's standards), and was limited in the amount of data that it could process. As a result, early GIS data structures were designed to minimize superfluous information and to maximize processing efficiency; only those attributes essential to the specific application were included. Likewise, since each application was unique, there was little need for data exchange or standardization of data formats.

The principal feature of interest in early GIS applications was an area. Areas were depicted, depending on the GIS technology, either as groups of contiguous rectangular picture elements (or pixels) in a grid (a raster GIS), or by irregularly shaped lines enclosing a polygon (a vector GIS). Raster GISs are relatively simple to understand and to program for area-based analysis, but locational accuracy and resolution is limited by the size of the pixels. Vector GISs, on the other hand, permit better display resolution and more accurate analyses, but require more complex mathematics for display and overlay analyses. As computer processing speed and data storage capacities increased, GIS technology generally moved away from raster and toward vector-based approaches.

## *2.2. Origins of transportation models*

Transportation network models also began during the 1960s, primarily to support planning for new highways and public transit systems in urban areas. The US Federal Highway Administration (FHWA) developed some of the first standardized computer-based software to implement transportation network models, and provided this software to state and local transportation planners in order to facilitate objective analyses of alternative transportation investments (Weiner, 1997). Like GISs, transportation network models were constrained by early computer technology. Consequently, data structures were designed to maximize processing efficiency and minimize superfluous information.

In transportation network models, the principal feature of interest was, and is, the network. A network is a mathematical concept that is used to describe and analyze the operation of a physical system such as a road, a river, a pipeline, or, even, a telephone system. The key attributes associated with a network are measures of capacity and impedance along each network segment, and locations on the network where trips can begin, end, or switch to another segment. Unlike the spatial data used in GISs, there is no requirement that analytical networks be geographically referenced.

### *2.3. Development of commercial software*

Throughout the 1970s and 1980s, GISs and transportation network models continued to evolve and mature along parallel but independent paths, with surprisingly little interaction between the two technologies. During this period, computer technology improved dramatically, becoming faster, more powerful, more affordable, and easier to use. Computer-based analyses that previously could be conducted only by federal government agencies, research institutions, or by purchasing time on a large commercial mainframe computer could now be performed by consulting firms, local government agencies, and non-government organizations using their own desktop computers. As computers became more common in the workplace, the market for application software also grew. While office application software such as word processing, spreadsheet, and database management packages enjoyed the largest commercial success, more specialized niche markets also developed for GIS software and transportation network models. These commercial software packages were typically adaptations of earlier computer programs, based on established concepts and data structures, with most of the enhancements directed toward improvements in processing speed, simplified graphical user interfaces, and output displays. Since the customers for GIS and transportation software represented separate and largely independent markets, there was little demand and therefore little commercial incentive to integrate the two applications into a single commercial software package.

### *2.4. TIGER and GISs*

Perhaps the most significant catalyst for introducing GIS technology to the transportation community in the USA was the development and widespread dissemination of the census TIGER/Line files in the early 1990s. In preparation for the 1990 decennial census, the US Bureau of the Census created a standardized geographic database of linear features including roads, railroads, rivers, shorelines, and governmental boundaries, covering the entire USA and all US territories where the census would be taken. This database was named TIGER, which was an acronym for the Topologically Integrated Geographically Encoded Referencing system. All of the various linear features represented in TIGER were collapsed into a single layer of intersecting lines. The intersecting lines formed the boundaries of polygons, which were then used by the US Bureau of the Census to build the geographic areas it used to collect and report demographic data for the 1990 census (e.g. census enumeration districts, census blocks, block groups, and tracts).

Although the US Bureau of the Census had previously developed geographic databases for the 1970 and 1980 decennial censuses, these earlier databases,

known as GBF/DIME files, were limited in coverage to only major metropolitan areas. Moreover, the GBF/DIME files were designed primarily for internal use; the bureau had no formal policies or mechanisms for public dissemination.

With TIGER, however, the US Bureau of the Census took on a much more proactive role in disseminating the geographic databases. Prior to the 1990 census, the bureau actively promoted TIGER to various potential user groups, including planners, demographers, and business marketing groups. TIGER/Line files were released using standardized, well-documented formats, on CD-ROM media, at very low cost. The bureau also worked with commercial GIS software vendors to develop programs for importing TIGER/Line files into the vendors' proprietary formats.

The US Bureau of the Census's active marketing and dissemination of TIGER was done in its own long-term self-interest. In developing TIGER, the bureau had created a *de facto* national geographic database standard. By promoting widespread use of this database, the bureau believed that it would be able to establish data-sharing partnerships, especially with state and local government agencies. These partnerships could increase the accuracy and reduce the costs to incorporate new geographic features like roads, and to update key attributes like address ranges.

In reality, the impact of TIGER was even greater than predicted by the US Bureau of the Census, but not exactly along the lines it had expected. The greatest impact of TIGER was in the field of "business geographics." By providing a national database of address ranges, tied to a geographic base map, TIGER made it relatively easy to pinpoint the locations of specific addresses on a map. A number of commercial start-up firms sprang up in the early 1990s to develop software for the automated matching of addresses with their locations on the TIGER-generated maps, and to create commercial spin-offs of the TIGER database containing enhanced and updated address information, improved geographic accuracy, and/or additional feature attributes. The creation of these proprietary versions of TIGER undermined, to some extent, the vision of the US Bureau of the Census of establishing data-sharing partnerships, but the success of TIGER as a spur for the widespread proliferation of GIS in the USA far exceeded all expectations.

## 2.5. *The Census Transportation Planning Package and GIS*

Another product of the 1990 US census that had an equal, if not greater, impact than TIGER on linking spatial and transportation data for urban transportation planning applications in the USA was the development and dissemination of the Census Transportation Planning Package (CTPP) (Transportation Research Board, 1995). The CTPP consists of special tabulations of journey-to-work questions asked

as part of the census long-form questionnaire. The census long form includes questions on workplace location, mode of transportation, and travel time from home to work, as well as residence location and demographic characteristics of the household. US federal privacy restrictions prohibit simultaneous release of detailed demographic and location data about a respondent in the same record. The US Bureau of the Census therefore prepares special summary tabulations of journey-to-work information at the census tract level of geographic detail, by residence location, workplace location, and aggregate flows by mode between residence and workplace.

Before the 1990 census, the US Bureau of the Census prepared these tabulations only upon request for a specific geographic area, and would charge the requesting organization the full costs for processing. With the 1990 census, the US Department of Transportation (USDOT), in cooperation with the American Association of State Highway and Transportation Officials (AASHTO), contracted with the US Bureau of the Census to prepare summary tabulations of its journey-to-work data for the entire USA. Summary tabulations were prepared at two levels of geographic resolution: (1) a statewide tabulation where origin and destination locations were aggregated to a county- or census-defined place; and (2) a metropolitan tabulation where origin and destination locations were aggregated to a Census tract or transportation analysis zone (TAZ). Metropolitan tabulations were produced for each urbanized area identified in the 1990 census with a population over 50 000.

The USDOT's Bureau of Transportation Statistics (BTS) disseminated the CTPP data free of charge on CD-ROM media. In addition to the summary tabulations, the CD-ROM included a simplified GIS software program that enabled the user to view the CTPP data geographically and produce simple thematic maps showing, for example, the percentage of work trips going to each destination zone. The geographic database used in the CTPP viewer software was the 1990 TIGER file.

The widespread distribution of the CTPP in the USA, especially among metropolitan transportation planners, not only demonstrated the value of census journey-to-work data for updating local demographic assumptions and forecasts, but also showed how geographically referenced data could be integrated with transportation network models using GIS technology. This provided the necessary impetus to move GIS and transportation network models from parallel to convergent paths of development.

### 3. Conceptual differences between GISs and transportation models

GISs and transportation models both use lines to represent linear transportation features such as roads or railroads. However, the lines are based on different conceptual models, with different data requirements and topological relationships.

Table 1 summarizes the key conceptual differences between GISs and transportation models that must be accommodated by additional attributes or through a translation procedure. These differences are discussed below.

### *3.1. GIS spatial objects and relationships*

GISs are a tool for displaying and analyzing spatial relationships between geo-spatial features. Typical spatial questions addressed by GISs include proximity (e.g. "How far apart are two features?"), adjacency (e.g. "Do two features share a common border or point?"), containment (e.g. "What features are located within a specified area?"), and connectivity (e.g. "What features are connected to one another?"). In vector-based GISs, geo-spatial features can be represented by one of three geometric objects:

- *Points.* A point is defined as a zero-dimensional spatial object that specifies location. The location of any point on the earth's surface can be described by a pair of geographic coordinates such as latitude and longitude. Points are used in GISs for several different purposes. They can represent the location of a physical feature such as a street sign; define the terminus of a linear feature such as a road segment; establish intermediate shape points for drawing a line; represent the centroid of an area feature such as a land parcel; or provide an anchor point for placing a label.

Table 1  
Differences between GISs and transportation network models

GISs	Transportation models
Primary analysis unit is an area	Primary analysis unit is a network
Lines used to delineate borders of polygons, which depict areas	Lines used to depict links, which are essential elements of a network
Planar topology required to support some spatial analysis	Non-planar topology reflects certain network features (e.g. overpasses)
Relative positional accuracy among layers is more important than absolute accuracy	Positional accuracy is not important, but link connectivity is critical
Required line attributes include the identities of start and end points, and areas to the left and right sides of the line	Required line attributes include the identities of start and end points, and permitted directions of flow
Attributes of areas are associated with polygon features	Attributes of areas are typically assigned to network nodes, which depict area centroids

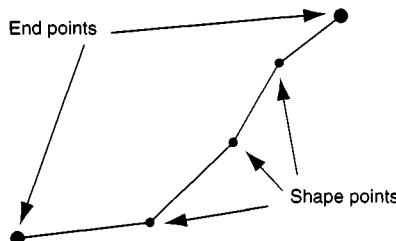


Figure 1. Lines represent linear geographic features in vector-based GISs.

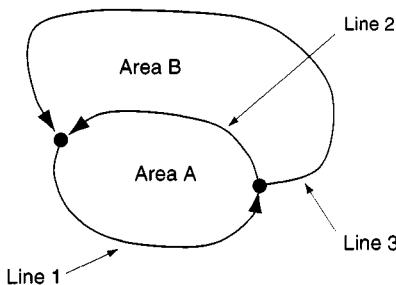


Figure 2. Polygons represent physical geographic features in vector-based GISs.

- **Lines.** A line is defined as a one-dimensional spatial object, which connects two or more points. A simple line segment can be fully described by identifying its start and end points. More complex linear features include multiple line segments, where adjacent segments share a common end point, as illustrated in Figure 1. The intermediate points are typically referred to as “shape points.” Lines are used in GISs either to represent linear geographic features such as roads, railroads or rivers, or to define the boundaries of area features.
- **Polygons.** A polygon is defined as a two-dimensional spatial object, formed by one or more connected lines, with closure (i.e. the end point of the last line forming the polygon boundary connects to the start point of the first line forming the boundary), as shown in Figure 2. A polygon can be fully described by identifying the lines that comprise it. This is typically done by including attributes with the line feature that identify the polygon located on the left and right side of the line, as it is drawn. In Figure 2, line 2 includes attributes that indicate that area A is on its left and area B is on its right. In a similar fashion, line 1 would indicate that area A is on its left, with no area on its right. Line 3 would indicate that area B is on its left, with no

area on its right. The lines that form polygon boundaries in a vector GIS typically represent some physical geographic feature, such as a river, coastline, administrative border, or segment of a transportation network. However, in many GIS applications, only the geographic shape of the line is important. Visual attributes of the line feature (e.g. whether it is a road, river, or county line) may be helpful for display purposes, but network properties such as impedance, flow or capacity measures are generally not relevant to area-based GIS analyses.

Area-based GISs also require an assumption of planar topology. Planar topology assumes that the lines used to define a set of area features lie on the same horizontal plane. For example, if two roads, a railroad track, and a river define an area feature, planar topology assumes that the roads, river, and railroad tracks all intersect on the same horizontal plane. There is no provision for any of the line features to cross another linear feature without intersecting it, as in the case of a bridge or overpass.

### *3.2. Network objects and relationships*

A network is a mathematical construct defined as an interconnected system of nodes and directed links, as defined below:

- *Node.* A node represents a decision point on a network. In transportation networks, nodes typically identify locations where travel can begin, end, or branch. In road networks, branch nodes most often correspond to intersections, where a traveler can choose one or more alternative routes. However, more complex, multi-modal networks can also be created in which branch nodes may represent an intermodal transfer point (e.g. a bus stop, rail station, or airport), where a traveler can switch from one travel mode to another.
- *Link.* A link represents a permitted path from one node to another. It has an explicit direction of travel, and at least one associated cost or impedance (e.g. distance, travel time, etc.). A link may also have an associated measure of capacity, indicating the maximum volume of flow allowed in a specified time period (e.g. vehicles per hour on a road, or gallons per minute through a pipe). A link has no explicit shape; it is usually depicted graphically as a straight-line vector between two nodes, as illustrated in Figure 3.

Although networks are often used to depict geographic features such as roads or rivers, they can also be used to model any system characterized by decision points and branching alternative paths. Examples of non-geographic networks include telephone exchanges, network circuit boards, manufacturing processes, or

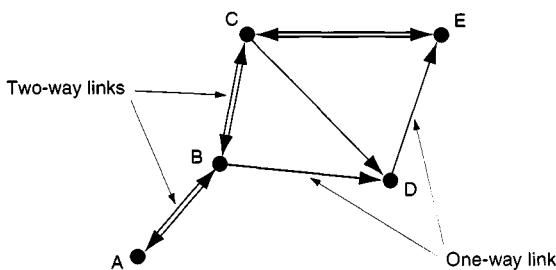


Figure 3. Links and nodes in a network.

even human thought processes. Consequently, unlike a geo-spatial layer used in GISs, a network does not require an explicit geographic location.

The primary topological property required in a network is connectivity. Each node in a network is an endpoint to one or more links. Network connectivity does not require an assumption of planar topology. One network link may cross another link without intersecting it. This allows, for example, the depiction of highway overpasses in which two roads cross but no valid path exists between them.

### *3.3. Translating between linear spatial objects and networks*

Points and lines can depict both linear spatial objects and network objects. As discussed above, however, GISs and network models require different attributes and assumptions. In order for the points and lines used in a GIS to be suitable as links and nodes in a network model, the attributes discussed below either must be included or added as part of a translation procedure.

#### *Permitted travel direction*

A line feature used in a GIS has implicit start and end points, which are used in drawing the line (i.e. by connecting a sequence of shape points), and in orienting the line to specify its left and right side for identifying polygons. This topological direction is independent of the direction of permitted flow in a network model. The direction of permitted flow on a link may be either the same as, or opposite to, the topological direction of the line, or flow may be permitted in both directions (e.g. where a two-lane road is depicted by a single line in a GIS).

In order to account for network flow direction, most network translation procedures recognize and/or create a link attribute that identifies the direction of permitted flow relative to the topological direction of the geographic line from

which the link was derived. For example, such a field could permit one of three values: 1 if the permitted flow is the same as the topological direction, -1 if the permitted flow is in the opposite direction, and 0 if permitted flow is in both directions.

### *Link impedances*

Network routing algorithms use link impedances to compute minimum paths through a network. Impedance measures for transportation networks may include distance, travel times, or some generalized composite cost. Line features used in a GIS typically include a length measurement, which may serve as a default impedance. However, route choices are more typically based on minimizing travel times, or on some trade-off between travel times and costs such as tolls. Moreover, travel time measures are functions not only of distance but also of facility type (e.g. speeds are typically higher on a limited-access freeway than on an arterial road with multiple signalized intersections). Therefore, additional impedance measures are often included in the line feature database, or are created during the GIS-to-network translation process.

### *Turn restrictions*

Network nodes represent points where two or more links intersect. In the absence of any additional information, movement is permitted from any link entering the node to any other link. In reality, however, movements between certain links may be constrained, or even impossible. For example, in a road network, it may take more time, on average, to turn left at a busy intersection, than to turn right or to proceed straight ahead (if driving on the right, as in the USA). Some turns may be prohibited by law (e.g. a "no left turn" sign), or may be physically impossible (e.g. where the node represents an overpass/underpass with no physical connection between the two roadways).

Some turn prohibitions can be handled by relaxing the assumption of planar topology in the geographic line database (i.e. allowing lines features to cross over each other without having any end points in common). Not all turn restrictions can be handled this way, however. Therefore, most GIS-to-network translation procedures also include a turn penalty table. The turn penalty table assigns an impedance measure to specific link-to-link movements at an intersection. Turn penalty tables are generally constructed so that they apply on an "exception" basis. In other words, only restricted turn movements are included in the table; all other turns are assumed to be permitted and without penalty. Turn penalties are incorporated into the network routing algorithms as additions to the link impedances, and should therefore be consistent in terms of measurement units (e.g. if impedances are measured in minutes, turn penalties should also be in minutes).

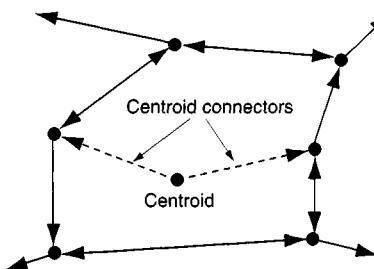


Figure 4. Centroid connectors in a network.

### *Capacity*

Measures of link capacity are needed primarily when capacity restrained assignment procedures are used. Highway capacity is measured in terms of vehicles per time period (e.g. passenger car equivalents per hour). Highway capacity depends on a number of factors, including the number of travel lanes, type of facility, terrain, vehicle mix, etc. The *Highway Capacity Manual* (Transportation Research Board, 2000) includes formulae for estimating highway capacities for any type of road in the USA.

As vehicle volumes approach (or even exceed) the capacity of a link, the impedance of the link will increase, making it less attractive relative to other links. A capacity-restrained assignment procedure loads traffic on the network using minimum impedance paths and then recalculates impedances based on the ratios of link volumes to link capacity. Traffic is then reassigned based on these revised link impedances. The procedure is repeated until there is little or no change in traffic or link impedance between successive iterations. This is the final “equilibrium” assignment.

### *Centroid connectors*

In transportation network models, all trips must enter or leave the network at a node. These entry and exit points typically represent small geographic areas (i.e. TAZs), from which trips originate or end.

As illustrated in Figure 4, centroid connectors are artificial links that connect each centroid to one or more physical nodes on the transportation network. Like real network links, centroid connectors have an associated impedance, which typically represents the average time to travel on local streets within a TAZ to reach a network link. Centroid connectors may also include special turn restrictions to prohibit traffic from taking a “short cut” through a TAZ rather than staying on the network.

Because centroid connectors represent a composite of local streets within a TAZ, they have no corresponding line features in a geographical database, and must therefore be created as part of the GIS-to-network translation procedure.

#### 4. Other transportation data structures

Beyond the conceptual differences in geo-spatial objects used by GISs and transportation networks, transportation applications also use several other data structures that are not common to traditional GIS applications. These structures are discussed below.

##### 4.1. *Routes*

A route is a complex linear object comprising a directed sequence of line segments or network links that share a common name or identifier. Examples of routes used in transportation include:

- signed highway routes (e.g. Interstate 95);
- named streets (e.g. Watling Street);
- urban bus routes;
- rail transit lines (e.g. Metro Orange Line).

As a geo-spatial feature, a route has the same geometry as the line segments that comprise it, but may have its own unique set of attributes. For example, the geographic shape of an urban bus route is defined by a sequence of street segments over which the bus travels. However, the bus route may also include attributes related to bus headways, passenger boardings and alightings, and bus stop locations that are unrelated to the street segments.

Different routes can also share one or more common line segments. As an example, Figure 5 illustrates a typical urban street network, with three bus routes. Bus routes 1 and 2 share line segment 203, and bus routes 2 and 3 share line segment 23.

Routes are typically represented in a GIS by identifying the sequence of line segments that make up the route in the order that they are traversed. In the example shown in Figure 5, the three bus routes would be defined by the following sequence of line segments:

- route 1 – 201, 202, 203, 33, 34, 35;
- route 2 – 401, 402, 24, 23, 203, 204, 205;
- route 3 – 21, 22, 23, 303, 304, 43, 42, 41.

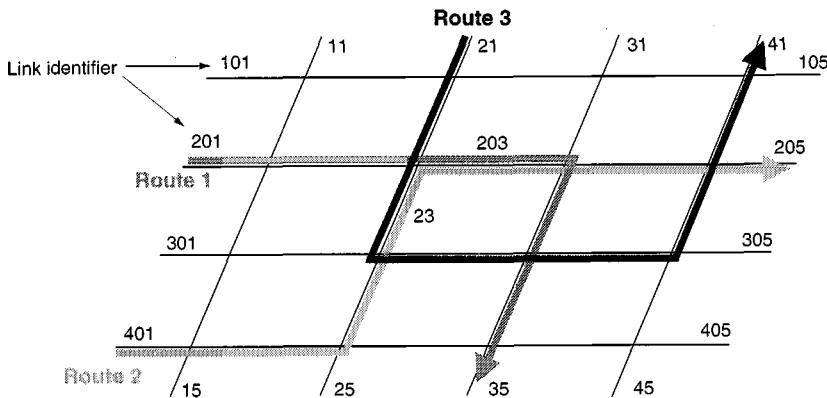


Figure 5. An urban street network with three bus routes (shaded lines).

#### 4.2. Linear referencing

Because geo-spatial features can be located anywhere on the earth's surface, they require two measurements (e.g. latitude and longitude) in order to uniquely specify their location. Features on a transportation network, on the other hand, require only one measurement – the linear distance from a known reference point on the route. The process by which features are located using linear measurements is known as linear referencing.

Transportation agencies used linear referencing to identify the location of road attributes (e.g. pavement condition), road features (e.g. bridges and signs), or events (e.g. vehicle crashes) long before they began using GISs. Most of the legacy databases maintained by transportation agencies use some type of linear referencing method to specify location. One of the key issues faced by state transportation agencies in adopting GISs was finding a way to match various linear referenced legacy databases to a geo-spatial road network (O'Neill and Harper, 1999).

The following information is needed to locate a feature using linear referencing:

- a route or line segment identifier;
- a measured distance from either:
  - the beginning of the route or line segment (the “milepoint”); or
  - a well-defined reference point along the route.

Figure 6 illustrates the difference between a linear referencing system (LRS) based on milepoints and a reference point system. Under a reference point system, any prominent feature can be specified as a reference point. Other points

of interest are measured relative to the specified reference point. In Figure 6, the bridge is selected as a reference point (RP1). The crash is 2.10 miles from the bridge, but simply providing a distance is insufficient because the crash could be located either east or west of the bridge yet still be on Route 10. Therefore, both a distance and direction must be specified using a reference point system.

A milepoint is a special case of a reference point. Under a milepoint system, the start of a route (e.g. Route 10) is assigned a milepoint value of 0.0. All other points of interest along the route are measured relative to the start of the route. In Figure 6, for example, the bridge is located 2.50 miles from the start of Route 10, the crash scene is 4.6 miles, and the end of the route is 5.4 miles. There is no need to specify direction in a milepoint system, because all measurements are made relative to the beginning of the route itself.

Most commercial GIS software can create a milepoint-based LRS as an extension of a route feature by adding beginning and ending milepoint values to each line segment associated with the route. Table 2 shows how such a route-segment table with linear referencing might look.

The route identified in Table 2 refers to bus route 1 in Figure 5. Each line segment associated with this route is assigned a start and an end milepoint value, starting at milepoint 0.00, where route 1 begins, and increasing to milepoint 2.75 where the route ends. Any feature along route 1 can be uniquely located with respect to a specific line segment, and can be further located along the segment

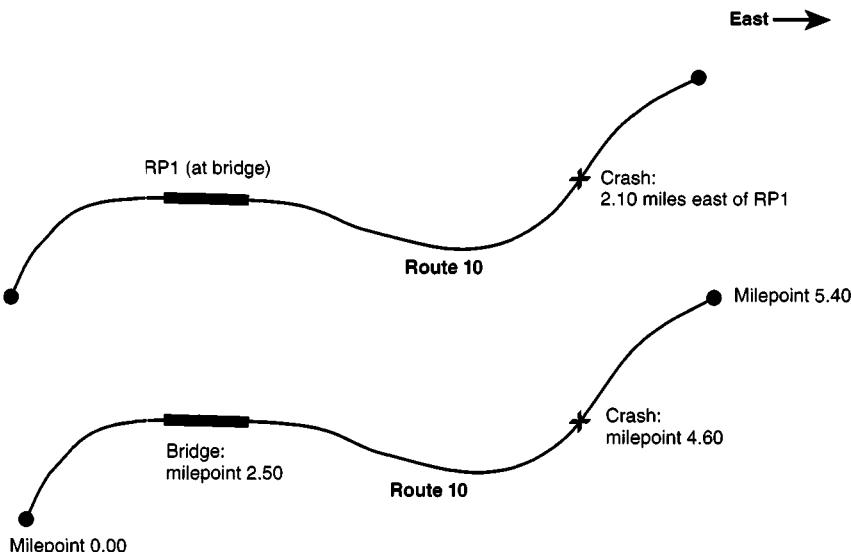


Figure 6. The difference between an LRS based on milepoints and a reference point system.

**Table 2**  
**Route-segment table with linear referencing**

Route	Line segment	Start milepoint	End milepoint
1	201	0.00	0.40
1	202	0.40	0.75
1	203	0.75	1.20
1	33	1.20	1.70
1	34	1.70	2.25
1	35	2.25	2.75

		Destination TAZ			
		1	2	3	4
Origin TAZ	1	0.0	3.6	7.9	5.4
	2	3.8	0.0	5.1	1.7
	3	8.1	4.8	0.0	2.1
	4	5.7	1.9	2.2	0.0

Figure 7. Example of a matrix using TAZ data.

through linear interpolation (e.g. milepoint 1.35 is located 30% of the way along the length of segment 33, measured from its intersection with segment 203).

#### 4.3. Matrices

A matrix is a data structure used extensively in transportation planning applications to store information about interactions between two locations. A matrix consists of rows and columns, as shown in Figure 7. Each row or column represents a specific location (e.g. a TAZ or network node), while the value in each cell represents some measure of interaction between the locations associated with the intersecting row and column. In the example shown in Figure 7, there are four TAZs, represented by rows and columns numbered 1 through 4. The cell entries represent the average travel time (expressed in minutes) between each origin and destination TAZ. The diagonal values, which represent travel from a TAZ to itself, have values of zero.

Examples of matrices used in transportation analyses include:

- trip tables – the number of trips made between two locations;
- travel impedance matrices – cumulative measures of impedance, or separation between two locations (e.g. distance, travel time, or generalized cost).

Even though matrices are often associated with geo-spatial features, they are not geo-spatial features themselves. Consequently, most conventional GIS software packages have no capabilities for displaying, manipulating, or integrating matrix data with geo-spatial features. Procedures needed to integrate matrix and geo-spatial data include:

- *Matrix creation.* The dimensions (rows and columns) of a matrix from a selected set of geo-spatial features (e.g. TAZ boundaries or network nodes) are defined. Each matrix row and column retains an identifier linking it back to the geo-spatial feature that it represents.
- *Matrix filling.* Global values are assigned to matrix cells; cells are filled using a formula; values are imported from external databases; and cumulative totals are extracted from network path-building procedures.
- *Matrix operations.* Matrices are added, subtracted, multiplied, and divided, and transposed (i.e. rows and columns are switched); and row marginals are computed (cell entries added across all columns for each row), as are column marginals.
- *Linking matrices to geo-spatial features.* Attributes are created from matrix rows, columns or cells, and then joined to geo-spatial features based on the feature identifier.
- *Creating desire lines.* “Desire lines” are straight-line segments connecting two geo-spatial features that can be used to display measures of interaction between the two features. These measures are populated by entries taken from the cells of the matrices.

#### 4.4. Dynamic spatial objects

Transportation objects are more spatially and temporally dynamic than most of the geo-spatial features typically analyzed by GISs. Much of transportation analysis deals with the movements of vehicles, goods, and people over links in a transportation system. The locations of these objects change over very short time intervals (i.e. minutes), and these changes are more often the subject of interest than the location of the object at a specific point in time. Even relatively static features, such as roadways, may change characteristics over a relatively short time period (e.g. reversible, high-occupancy vehicle (HOV) lanes, may change permitted direction of traffic flow several times over a 24 h period).

Until recently, neither GISs nor transportation network models dealt very effectively with display or analysis of dynamic spatial objects. GIS displays or changes over time were typically limited to “thematic snapshots,” showing values of a particular attribute (e.g. population density or traffic volumes) at two or more points in time. Similarly, transportation network models generally partitioned

daily traffic patterns into two or more representative time periods (e.g. peak versus off-peak), and conducted separate model runs for each period.

Better understanding of transportation phenomena and travel behavior requires improved tools for spatial-temporal analysis and display. Some of these tools are now being incorporated into both GIS software and transportation models. These include:

- *Vehicle-tracking algorithms.* The location of any object can be represented as a point feature in a GIS. Using Global Positioning System (GPS) receivers mounted in moving vehicles, it is possible to record the locations of these vehicles as they traverse the transportation network. These locations and the time they were recorded can either be transmitted directly to the GIS, or saved in a geo-spatial file. The point features representing the vehicle positions can be overlaid on a network layer to show vehicle movements as they occur, or using a playback procedure to simulate movements in compressed time. An important consideration in displaying vehicle movements is the relative positional accuracy of both the kinematic GPS measurement and the transportation network it is moving over. GIS software may include algorithms to “snap” the point locations to the nearest link in order to improve its display.
- *Microsimulation tools.* A new generation of transportation network tools is using simulation to analyze the movements of individual travelers and vehicles rather than averaging volumes over a specified time period (Shunk, 1994). Microsimulation enables transportation analysts to investigate dynamic phenomena such as queuing behavior at intersections and on-ramps, to determine the impacts of parking and incident management policies on traffic flow, and even to track the daily trip patterns of individual households. Visualization of microsimulation results is similar to vehicle-tracking displays, but does not require the same degree of positional accuracy for the underlying network.
- *Spatial-temporal data research.* Basic research into the data requirements and new tools needed to incorporate dynamic spatial objects into GISs was recently undertaken (Adams et al., 2001). The study identified the basic conceptual properties of dynamic spatial objects, the operations that are needed to effectively represent them in GISs, and proposed a new data model structure to meet these needs.

## 5. Conclusions

GIS technology offers significant benefits in transportation applications. These benefits go beyond improvements in data visualization and presentation. They

also include better methods for data management and data integration using geo-spatial queries and spatial analysis techniques. However, only when developers and users understand the strengths and limitations of both technologies will the full potential of GISs in transportation be realized.

## References

- Adams, T.M., N.A. Koncz and A.P. Vonderohe (2001) *Guidelines for the implementation of multimodal transportation location referencing systems*, NCHRP Report 460. Washington, DC: National Academy Press.
- Burrough, P.A. (1990) *Principles of geographical information systems for land resources assessment*. New York: Oxford University Press.
- O'Neill, W. and E. Harper (1999) *Resource guide on the implementation of linear referencing systems in GIS*. Washington, DC: Bureau of Transportation Statistics.
- Shunk, G.A. (1994) *TRANSIMS project description*. Arlington: Texas Transportation Institute.
- Transportation Research Board (1995) *Decennial census data for transportation planning. Conference proceedings 4*. Washington, DC: TRB.
- Transportation Research Board (2000) *Highway capacity manual 2000*. Washington, DC: TRB.
- Weiner, E. (1997) *Urban transportation planning in the united states: an historical overview*. Washington, DC: US Department of Transportation.

***Part 5***

**GIS APPLICATIONS**

## THE ROLE OF GIS IN LAND USE AND TRANSPORT PLANNING

HOWARD L. SLAVIN

*Caliper, Newton, MA*

### 1. Introduction

Geographic information systems (GIS) have become an essential technology for land use and transport planning. The last decade of the twentieth century brought rapid evolution in all forms of computing and in the evolution of GIS, culminating in the widespread availability of GIS software that is powerful, low cost, and runs on inexpensive desktop computers and servers. While GIS are not as widespread as office productivity software, they have been broadly adopted by planning entities and the consultants who serve them.

Maps have always occupied a central role in planning, and today's digital maps provide an unlimited range of map views and printed products that enrich everyone's understanding of existing and planned development, locations for new facilities, and accessibility provided by transport. Databases have always been desired by decision-makers, and GIS have made a powerful contribution to planning by making it possible to collect, create, and manage vital data needed for many forms of planning. Query and visualization of geographic information rather than formal analysis dominate most GIS applications, but in some areas, such as transport modeling, analysis is more common. GIS technology itself has been enabled by the collection of large amounts of spatial data, which has further stimulated applications. The increasing spatial resolution of available data is enlarging the quantity and quality of planning applications.

The focus in this chapter is on the use of GIS for data development, presentation, and, especially, modeling. The last topic raises some of the most interesting issues, and has great promise for improving our understanding of transport systems. Some familiarity with the concepts and uses of GIS is assumed, and readers may wish to refer to the first handbook in this series (Dueker and Ton, 2000) for additional background on GIS basics and GIS applications to transport. A cautionary note is that there is a fundamental danger in writing about the state of any software technology, as rapid change is characteristic of software, and the literature is soon out of date.

## **2. GIS in land use planning**

GIS are now commonly used to prepare and illustrate land use plans and, to a great extent, digital mapping has become the principal graphics tool of the urban and regional planner. Existing land uses, zoning maps showing permitted land uses, and planned developments are routinely created and disseminated by planners both in hard copy and in digital form.

### *2.1. Data development, presentation, and access*

GIS and related technologies have facilitated the collection of a substantial amount of geographically referenced data and placed it at the fingertips of planners. Data layers typically encountered in a municipal or regional GIS installation include road centerlines, tax parcels, building footprints, zoning districts, census jurisdictions, and political districts. Elevation grids and digital orthophotography and/or satellite imagery are also commonly available for entities that are large enough or have sufficient wherewithal to have funded GIS activities. Remote-sensing procedures that process various forms of satellite imagery and aerial photography have been developed to classify land cover and identify changing land uses based upon the electromagnetic signatures they emit or reflect. However, use of remote sensing is not yet in widespread use by municipalities.

All of the available geographic and attribute data can be easily displayed and used to create a wide variety of attractive and informative printed and electronic maps. As shown in Figure 1, the thematic map, in its many incarnations, provides color-coded maps that indicate the land use status or other attributes of land parcels.

If a region is not flat, there is no insurmountable mapping problem; current software makes it easy to drape maps and imagery on surfaces so that three-dimensional representations of elevations are readily available. A GIS also makes it possible to develop representations of the built landscape that sit atop the land surface. Other visualization tools for virtual reality can take GIS input and transform it into a cityscape, an example of which is shown in Figure 2.

### *2.2. Data access*

Of equal importance to its use for data presentation, GIS technology has become a dominant information transfer mechanism for social and economic data. GIS are the principal means of access to census data in most countries, and have greatly contributed to the collection of this information. The landmark TIGER



Figure 1. Land use and parcel zoning.



Figure 2. Tax parcels and extruded building footprints on a draped aerial photograph.

effort in the USA not only improved the efficiency of the census, but also created the public domain data files with which the census data could be mapped by virtually anyone with a personal computer. This was accomplished despite the fact that the geographic accuracy of the TIGER files was poor in many areas and insufficient for many applications.

The provision of census data in GIS-compatible formats did more than simply provide access to specific numbers; it made the data comprehensible and had the potential to reveal important spatial patterns. For example, Figure 3 illustrates the geographic distribution of owner-occupied versus rental housing units and the positive correlation of owner-occupied housing with the median number of rooms per housing unit for Washington, DC, and its environs.

Many other government agencies, particularly in the USA (e.g. the Department of Housing and Urban Development, the Environmental Protection Agency, and the Department of Transportation), followed the example of the US Bureau of the Census by publishing their own data in GIS formats, resulting in a great expansion of data sets that are valuable for planning. One example is that of the US Department of Housing and Urban Development's Community 2020 Initiative and eMap service. Web-based systems in the USA provide good access to information, especially if one needs specific data items in small quantities. (See, for example, American Factfinder provided by the US Bureau of the Census.)

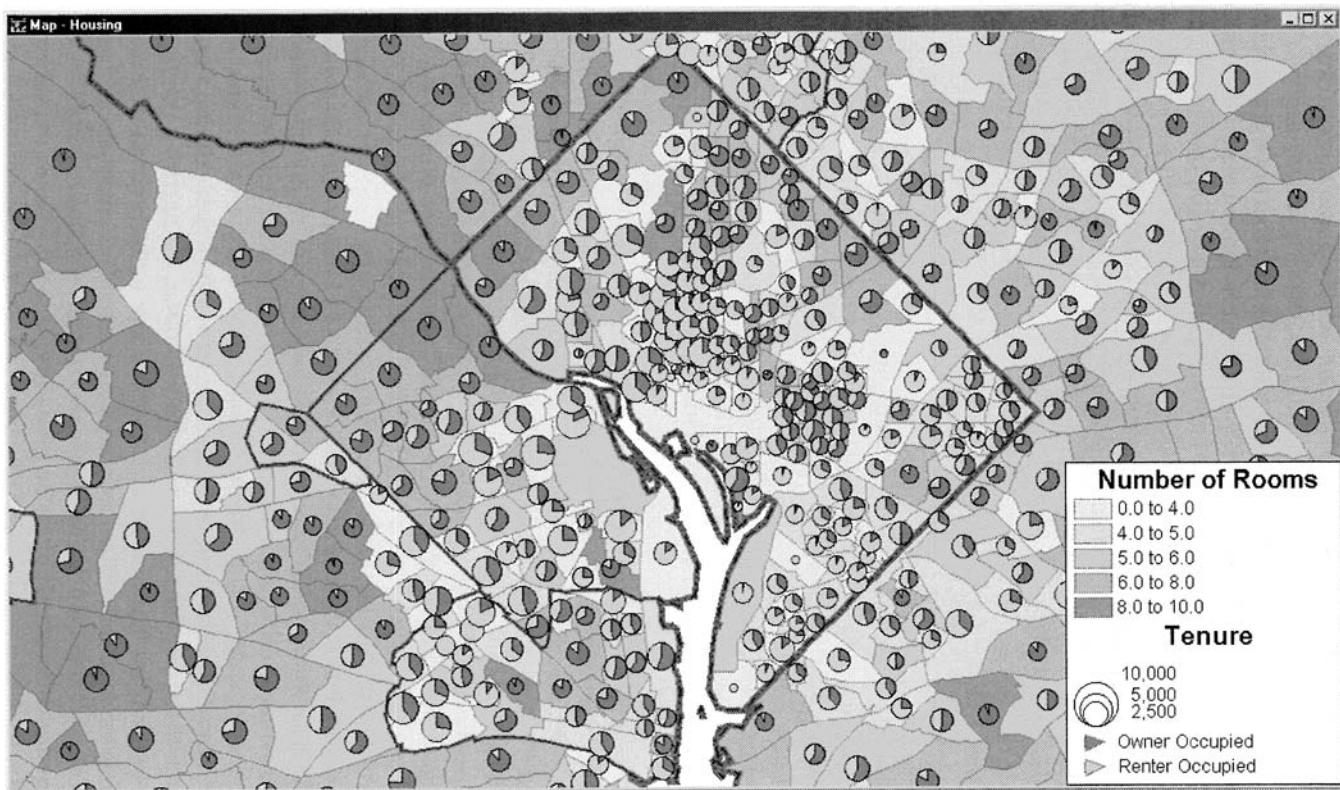


Figure 3. Housing unit tenure and median number of rooms.

There are also specific tools and data CDs that have been developed to satisfy the needs of those who require the voluminous detail that is published.

The acceleration of public information available has in a considerable measure come from maturation of GIS and their ability to link to virtually any relational database. Earlier GIS imposed proprietary data models that incorporated both positional and attribute data in the same data files. This made it much more difficult to utilize information that resided in other databases. Now data can be maintained in virtually any form and by other departments in government and still be accessed easily with GIS software.

### *2.3. Urban information systems and urban analysis*

Despite this wealth of technology, extant urban information systems are very limited, and the role of GIS in planning is predominantly associated with recording what has been planned or developed rather than in aiding the planning process. More often than not, local governments are departmentally balkanized, and their GIS departments are fully occupied with maintaining basic information on real estate parcels and to a lesser degree on infrastructure as reflected in road centerlines and sewer locations. This leaves little time to support ongoing planning.

Another impediment to the proactive use of GIS in planning derives from the fact that without application development, a GIS may be unsuitable or too complex for anyone to use. Even those organizations large enough to have a GIS department may lack strong skills in software design and implementation. Also, once developed, even simple applications are rendered obsolete by changing operating systems and frequent changes in the scripting languages used by some of the leading GIS packages. Consequently, GIS technology becomes expensive and largely unproductive. This is not a phenomenon unique to GIS, and is a well-known problem in information technology.

There is also an absence of generally accepted techniques for urban analysis, and many of the methods that are well established in the technical literature may not be uppermost in the minds of urban planners. These methods have not been incorporated in commercial GIS offerings because software providers rarely supply features that are not in demand. Consequently, much of the potential for GIS applications in land use planning is untapped. In particular, GIS should be used to perform basic analysis as well as to monitor change and trends and help in understanding the dynamics of urban and regional development. If the potential for GIS-based analysis is largely unrealized, GIS may find greater application in descriptive or predictive models of land use, growth, and development.

### 3. GIS in land use modeling

GIS are an enabling technology for land use modeling, and typically are either linked to or is used to feed data to most current land use models. The useful connection between GIS and land use models was recognized early on. For example, one of the first uses of the SYMAP computer mapping software was to visualize the outputs of land use simulations.

If that connection was lost for a time before the personal computer era, it was due more to the expense and difficulty of producing computer graphics than to any fundamental lack of relevance. Early GIS software was too daunting for most researchers to use, causing it to be underutilized in modeling. Consequently, as in transport planning software, GIS use was bypassed for a while. As municipal GIS implementation proceeded, land use modelers could not wait to obtain the data they held. By 1995, attendees at a conference on land use models universally acknowledged the value of GIS data and GIS technology in creating land use models.

Although none are in widespread use, there are many land use models that have been developed. Thirteen are described by Wegener (1995). Many of the more recent models have been linked to some form of GIS for the gathering of inputs or the presentation of results. The California Urban Futures Model (Landis, 1994) is built upon a GIS platform, and predicts housing development for sites; however, it does not treat transportation as a determinant of land use. In models that do include transport, both zonal (polygonal) and grid cell structures have been used, and historically most models were aggregate in nature, as were the prevalent transport forecasting models in use. In general, land use models have not been very detailed geographically, and tend to be coarser in detail than the transport models that are used. There are computational limits, and some data may not be available at more detailed geographic scales. Irrespective of scale, there are difficult issues attendant to the choice of zoning systems, all of which are accompanied by aggregation bias and other biases of unknown degree (Green and Flowerdew, 1996).

With the maturation of choice models, land use models are becoming more disaggregate and behavioral in nature, and will require and benefit from the greater geographic detail that GIS technology can provide. For example, the UrbanSim model development effort (described in more detail in Chapter 9) incorporates disaggregate data which is aggregated to a grid layer with square cells that are 150 m on a side. Parcel-based land use models can be envisioned, which may be the most natural level for modeling (Waddell et al., 1998).

GIS technology offers the prospects of providing great volumes of data for modeling from existing data systems and tools for obtaining and creating other data needed for land use models. Aerial photography and remote sensing can

provide measurements of variables that are important but are not captured in the municipal GIS. Buffering and polygon overlay can be used to create new variables.

A GIS has many facilities for managing spatial data of diverse types, and thus offers useful data management for the entities of interest in a land use model, be they areas, roads, households, residential dwellings, land use polygons, shopping centers, or employment sites. Any data of interest can be associated with these polygons, lines, or points.

The analytical uses of GIS should be fundamental to identifying and quantifying important relationships between transport and land use as well as other determinants of land use such as rents and economic growth. The ability to compute spatial statistics also can be used to identify correctable biases in model equations.

Having said this, most land use models do not exploit the power of GIS technology to a very great extent. At the heart of this is that, with but a few exceptions, land use models are designed to be linked to rather than integrated with a GIS. This makes using GIS procedures directly in land use models cumbersome or impossible. This may simply be a developmental stage, as was seen in transport modeling, an insufficiency or unsuitability of some GIS software, or it may reflect the reluctance of land use modelers to limit their work to a specific GIS platform.

What can be envisioned for future integrated models? Perhaps the most important difference would be the ability to use spatial queries and dynamically updated GIS calculations of land use area by type, adjacency, proximity, connectedness, accessibility, visibility, terrain, spatial autocorrelation, and overlay directly in the model components and equations. For example, in the micro-simulation approach based on choice models, GIS functionality can be used as an aid in composing choice sets and in characterizing alternatives.

The GIS database for an integrated model would include modeled futures as well as the past and base case measurements. Land use change itself could be computed and displayed graphically as well. A common interface for model query and a GIS would facilitate communication and understanding of the model forecasts. Of course, one can imagine a GIS-based land use model that is also fully integrated with a GIS-based travel-forecasting model. This has been done in a very preliminary and partial way in work undertaken on a successor to the STEP model (Harvey and Deakin, 1996; Caliper Corporation, 2002a,b). Many of the advantages of GIS for transport modeling also accrue to land use models; these are discussed in a subsequent section of this chapter.

While there is no insurmountable conceptual barrier to integrating GIS and land use models, GIS technology could also become more suitable for land use modeling by incorporating additional functionality that would support model development and application. Some of the desirable improvements are features of a GIS for transportation (GIS-T) for calculating accessibility and network travel times, but greater support for the time dimension would also be important as land

use models are inherently dynamic. Other features that could be added to a GIS would be support for model management in general and specific support for the spatial choice models that are used in the latest land use models.

Predictive land use modeling is a formidable endeavor, and involves many issues and difficulties that have nothing to do with GIS technology or transport (Lee, 1973). Nevertheless, a GIS makes it more feasible to implement models and should contribute to their greater effectiveness.

#### 4. GIS in transport planning

GIS usage in transportation planning lags behind that in land use planning but is catching up. In a sense this is more of a technology diffusion issue rather than one of solution availability. There is now well over a decade of experience in applying GIS to transportation planning. Stimulated in the USA, in particular, by the public domain TIGER street and census boundary files, there are scores of different types of transportation applications of GIS that have been successful. These would include facility and pavement inventories, emergency response systems, corridor alignment studies, project visualization, hazardous materials routing, snow plow and garbage collection optimization, fire station location, transit planning, freight analysis, and travel-demand modeling. GIS technology has also been granted a central role in the development of transportation information systems such as those that were mandated for congestion management and other applications.

Travel-demand modeling has always been the most computerized and computer-intensive operation in transportation planning. Thus, it is not surprising that GIS are playing an increasing role in demand forecasting and in the software with which they are implemented. In fact, the use of GIS in demand modeling is now almost universally accepted.

Much of the work on GIS-T applications has been pursued at the state or national level, and focused on issues that may not take on the same importance at the municipal or regional level, where data collection activities may be significantly less intimidating or make use of different concepts. Nevertheless, regional transportation planners are embracing GIS technology, and are now making use of it in many ways. At the same time, they are making demands on GIS-T that will lead to improvements in the future.

##### 4.1. An overview of GIS-T functionality

As commonly described, a GIS-T is a GIS that has additional capabilities for transportation. A GIS-T has extensions for handling transportation data objects

and procedures for applying transport analysis methods to be used on their own and as building blocks in more complex systems. Prevalent GIS-T programs have a full range of standard GIS functions for mapping and analysis. These include support for geo-relational databases, spatial indexing that makes it possible to handle very large databases without undue processing time, support for a wide range of map projections, flexible thematic mapping, surface analysis, and a broad set of geo-processing functions such as spatial queries and selection, buffering, and polygon overlay. Capabilities for building custom applications are also provided.

It is generally recognized that there are two species of GIS-T software (Waters, 1999). These are general-purpose GIS, such as ESRI's ArcInfo, and specialized GIS-T software, such as Caliper's TransCAD. The former has capabilities (extensions) for linear referencing and network analysis while the latter integrates transportation objects and analysis procedures directly. Based on the application at hand, there are arguments in favor of both approaches, and both types of software are often used together. For example, some organizations have used TransCAD as middleware to form a bridge between a general-purpose GIS and specialized planning software. Moreover, TransCAD is most commonly used in organizations that use ArcInfo as their institutional GIS.

If there were no advantages to a specialized GIS-T, none would exist or be used. While it may lack some other GIS functionality, a specialized system will have deeper functionality for transport. The presence of some isolated feature for transport is not really the important difference. Rather, it is whether or not there is a sufficiently complete system to perform the desired transportation analysis. Also, there may be a higher level of performance needed for certain operations than is found in a general purpose system.

Some of the advantages of a specialized GIS-T derive from the fact that its features are functionally driven/required and not simply designed and implemented for generic application. This raises the probability that the needed functionality will be present and not unduly complex. A specialized GIS-T is less complex for transportation professionals to understand and use for many reasons, not the least of which are use of a tailored user interface and familiar terminology. A specialized GIS-T may also be targeted more toward end use by transportation planners rather than a GIS department that fills their requests.

Of course, a specialized GIS-T implements generic functions for many transport applications, so it itself can be more generic and less efficient than some highly specialized application software. Inevitably, different packages will provide different solution capabilities and efficiencies. Perhaps most importantly, there is no longer any reason to use just one GIS or GIS-T. They can be inexpensive relative to professional staff costs, and, in the USA, most departments of transportation at the state level use two or three different packages.

### *Networks and path finding*

Network analysis is perhaps the single most important function of a GIS-T. This entails the ability to represent network topology for the purpose of finding the shortest paths in terms of time and distance on a road network. A network graph defines potential movements from place to place, including prohibited and permitted connections and the possible direction of movement on a link in terms of whether it is one-way in a particular direction or bidirectional. The network is used to find the shortest paths between points based on distance or time or some composite cost function. A GIS-T provides accurate measures of distance and possibly other attributes for the network. For example, it makes it possible to use road curvatures and grades in models. As shown in Figure 4, the grade of a truck route may be worthy of consideration in modeling. It can also be used to automatically create centroid connectors if these are needed. Another common function would be to compute isochrones or network bands reflecting the travel time from a location to all others, as shown in Figure 5.

In a GIS-T, networks are derived from line layers representing roads, rail lines, or other modes. While this can be straightforward, there are some subtleties to be considered. Multiple networks are needed to represent different time periods or different permissions such as truck routes or high-occupancy-vehicle (HOV) lanes. Also, it should not be necessary to build a completely different network in order to represent a different scenario.

A range of shortest path-finding capabilities is also a component of a GIS-T. It is extremely important that these be highly efficient, as some applications such as traffic assignments utilize many millions of shortest-path calculations. For many applications, the path finding should be done on congested networks, and reflect turning delays and ideally time of day effects (i.e., dynamic shortest paths). A GIS is also helpful in creating centroid connectors and building networks for other modes such as bicycling or walking.

### *Matrices*

Matrices are another key data structure for a GIS-T. Matrices hold flow tables, inter-zonal or inter-nodal distances and travel times, and are one of the most commonly used data objects used for computation in transport. The ability to spatially reference a matrix to points/nodes, line segments, or areas makes the matrix a particularly useful GIS construct. With spatial indexing, matrices can be labeled based on geography and aggregated to higher levels of geography. Another useful feature is viewing a submatrix that is based upon a spatial selection.

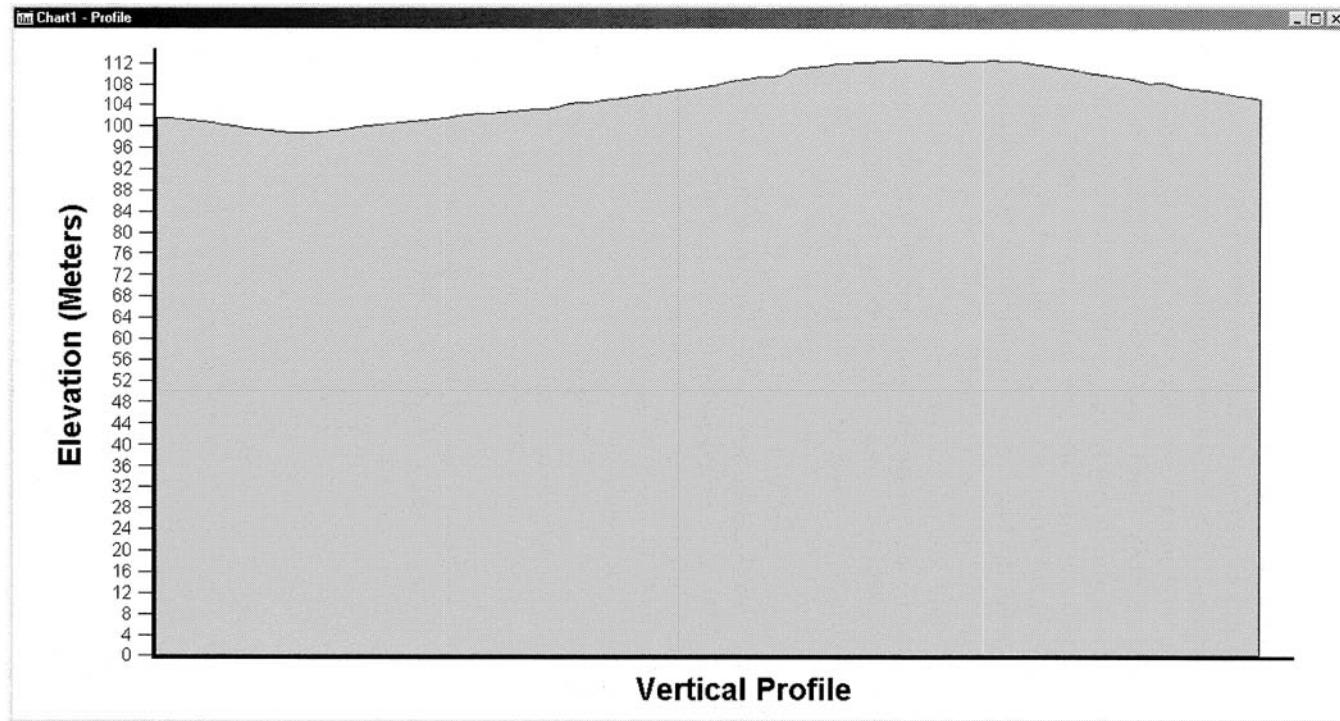


Figure 4. Vertical profile of a highway.

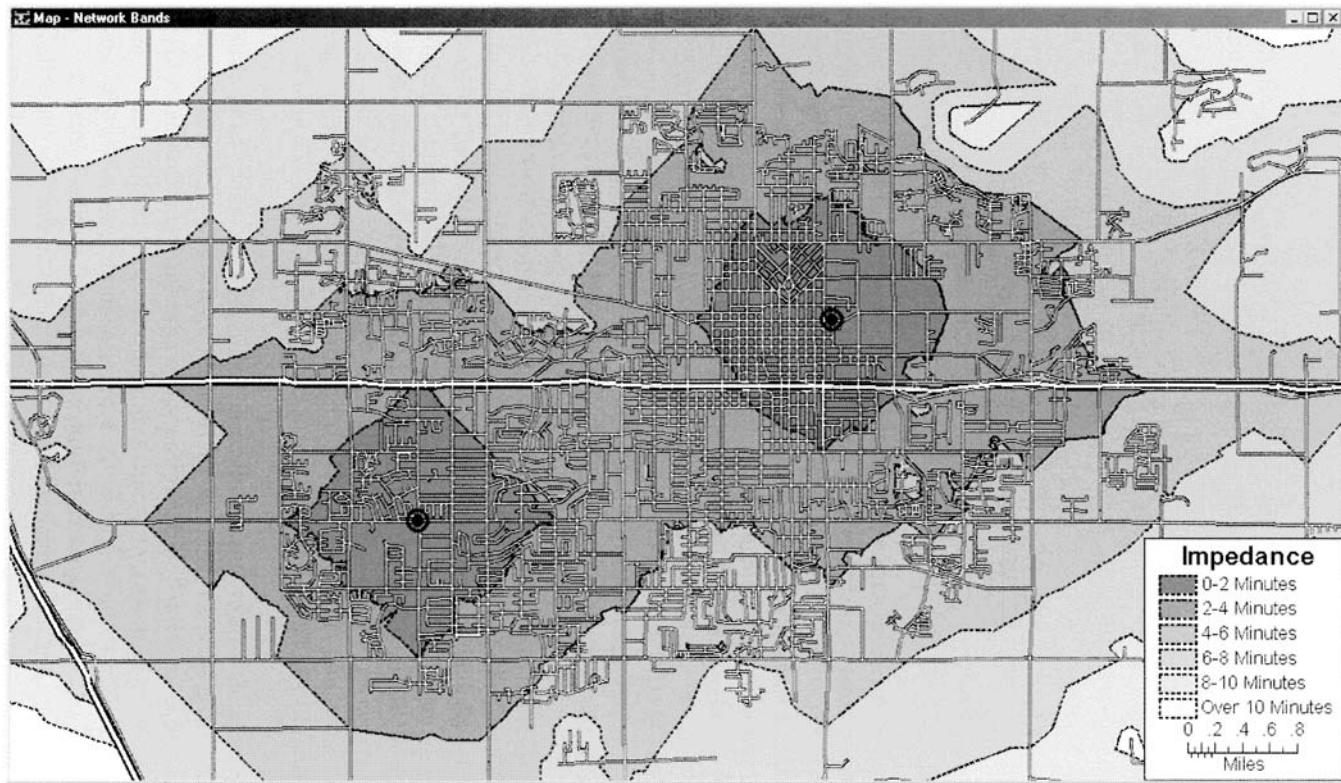


Figure 5. Network bands.

### *Routes*

A route is a directed series of links that connect an origin and a destination. In a GIS-T, it is very convenient to be able to store routes, edit them, and attach attributes to them. The last operation is important since some attributes are not link based but rather pertain to the overall route such as the headway or frequency for a bus route or the cost of shipping freight by truck over long distances. Repeating the route attribute as a field in the link database is inefficient. It is also important to maintain the relationship of a route to the geographic line layer on which it is based. Thus, when a road is moved, the route is modified. The functionality to build a network graph from routes is also useful.

### *Linear referencing and dynamic segmentation*

Some transportation data come in the form of locations specified by route and milepoint. In other words, the locations are specified by a linear distance on a route from a specific reference beginning point. Before the Global Positioning System (GPS) was created, linear referencing was a convenient way to specify locations along highways. For analysis, it was tempting but overly cumbersome to create segments any time that any roadway attribute changed. Dynamic segmentation overcomes this problem by conveniently creating the segments that are needed for a particular purpose.

A preoccupation with linear referencing and ameliorating its defects and removing its ambiguities has marked GIS-T research for more than a decade. Many feel that without standards and improvement to the underlying data model, data sharing will remain problematic. At the regional or metropolitan scale, linear referencing is less important for obtaining or maintaining road data although it certainly facilitates exchanges between metropolitan areas and states. Transit routes are often characterized as linear referenced entities where the stops and other significant points are located in terms of their distance from the beginning of a route.

### *Graphics and visualization*

GIS are characterized by superior map graphics, and provide many alternatives for visualizing geographic data. Available systems have some integrated charting capabilities that can also be used for data exploration and graphics. A GIS-T will have facilities for visualizing transport data of various types. Typically, these will include the ability to map flow data in ways that are not always handled in a standard GIS or mapping program. One example is shown in Figure 6.

Regional transportation authorities typically have basic GIS capabilities, and exercise them for mapping. General use of a GIS in transport planning typically

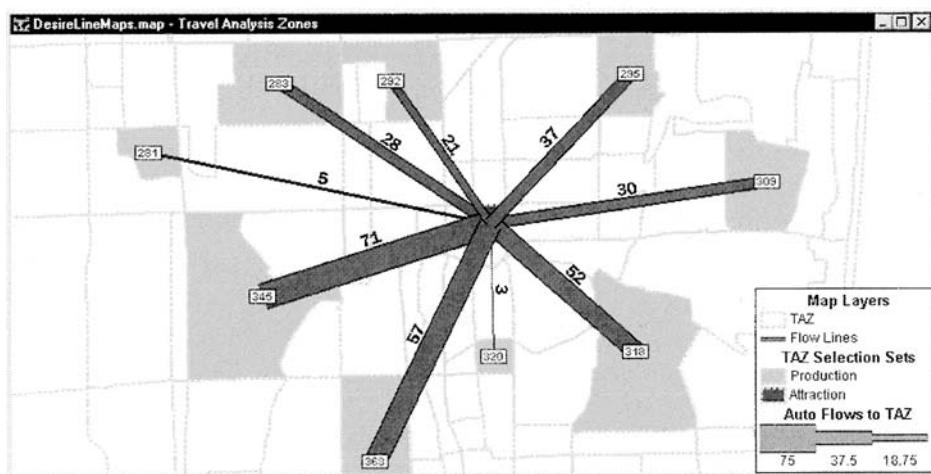


Figure 6. Desire lines showing flows.

takes on a facilities orientation that focuses on the roadway inventory, centerline file, or both. Depending upon the level of government, the same file may be used for many other purposes, such as emergency response or school district boundary alignment.

In the USA, metropolitan planning organizations nearly universally use GIS to access census data pertinent to transport planning. This is done at the traffic analysis zone level as well as at other geographic scales. In addition, GIS are used by transport planners to map and present data on traffic counts, accident locations, and, in some instances, the locations of traffic signs and signals.

GIS maps are used to illustrate reports and to provide graphics for the regional transportation plan. There is a growing trend to geocoding transportation improvement projects, and extensive mapping is used to illustrate projects that will be undertaken or are underway. GIS-derived graphics have become a mainstay of public meetings in which projects and alternatives are presented and discussed.

Transit planning is another major application of GIS-T technology, which is used to prepare transit maps for publications, to illustrate route alignments and transit service areas, and to estimate the population that is served. A GIS-T can display ridership and revenue, as well as be used to find the best way to travel from point to point on the transit system. For the service planner, a GIS-T can be used to compute the transit accessibility of different destinations from any origin and finds a similar use in planning and evaluating transit for the elderly and disabled. Even the *US Highway Capacity Manual* now endorses the use of GIS for computing transit accessibility.

GIS technology is raising the standards for the data files that are used in transport planning and modeling, as well as enabling the use of better data. Perhaps nowhere is the application of GIS more prominent now than in travel forecasting.

## 5. GIS in travel-demand modeling

GIS technology is more of a factor in travel-demand modeling than in any other facet of transportation planning. GIS have been embraced by model builders, as they have opened new vistas in data availability and data preparation for models. Others have seen the potential in fuller use of GIS in the modeling process as integrated GIS and transportation modeling software has evolved. With the introduction of new methods, GIS-T will continue to transform the way that travel-demand modeling is conceptualized and performed.

### 5.1. *GIS-T use in modeling: the linkage-integration debate*

Early in the development of PC-based GIS, it was recognized that this technology offered great potential to support travel-demand modeling. To a certain extent, this was accelerated by the use of GIS in the 1990 US census and in processing and disseminating the results to transportation planning agencies. Specialized software was developed for the US Bureau of Transportation Statistics to disseminate the journey-to-work data, and virtually all GIS provided some form of access to the tract, block group, and transportation analysis zone (TAZ) level socio-demographic data.

The initial roles accorded to GIS in modeling were the preparation of inputs and the visualization of outputs. These roles could be accomplished by having the ability to exchange data between planning software and GIS software. To many, this suggested the approach of linking the two types of software.

The alternative view of integrating GIS and transport models in the same software package was first and still best exemplified by TransCAD. Introduced in 1988, TransCAD was unique in that it was a combined GIS-T and modeling system, although full modeling capabilities were not added until 1993 and beyond. Nevertheless, the recognition that a GIS-T could be a suitable platform for a travel-demand forecasting system was initially a leap of faith that not everyone embraced. Yet, it was not so farfetched, particularly if it is recognized that mainframe UTPS was an early, if not the first, GIS-T application.

Linking planning software and GIS became a popular activity in the early 1990s (Anderson and Souleyrette, 1996). This was done with all of the planning

packages and most GIS software, including ArcInfo, MapInfo, Maptitude, and TransCAD. This is still being done in the USA and elsewhere.

Gradually, some of the unresolved difficulties of linking two disparate types of software together became more apparent. The first was that cumbersome import/export operations needed to be performed over and over again with each additional scenario that was to be analyzed. Second, the linkages were only partial. There would be data in one system that could not find a home in the other. A matrix in a dedicated planning package, for example, would have no direct counterpart in a GIS. The shape of the links in a GIS could not be displayed in the planning package. Third, the linkages were not always permanent. Changes to the GIS scripting language would break the combined application's interoperability.

Users of these combined systems had a steep learning curve, in that they needed to be well versed in two packages in order to use both with fluidity. Often a third or even a fourth piece of software was needed to assist in the import/export process or to perform some statistical analysis. This introduced additional complexity, and was particularly cumbersome when data sets were large.

Perhaps the greatest weakness is that data edits in one piece of software were not automatically reflected in the other software, leading to incompatibilities and inconsistencies which were further exacerbated by incompatible data structures and the lack of a common programming language. The latter was a barrier to effective applications development, and the lack of access to source code prevented minor modifications to each piece of software that would provide better support for demand modeling.

In contrast, there were many advantages to an integrated system. It permitted the fullest and most convenient use of a GIS in the modeling process and offered a higher degree of functionality and data integrity than could be achieved in linking software packages. All of the GIS functionality was provided in a form that could be used directly in modeling without import/export operations. It also offered the opportunity to build models at all geographic scales, thus providing a significant option for state-wide and international models. Application development with a single scripting language could incorporate both modeling and GIS procedures.

Potentially to its detriment early on, seamless integration of GIS and demand models was known to require additional coding of software, to have some potential computational overhead, and to involve some novel data structure issues in GIS-T. None of these barriers have proven to be insurmountable.

The possibility of linking a GIS-based travel-demand modeling package to a general-purpose institutional GIS was not initially considered by participants in the linkage-integration debate, but it has become commonplace, and may well prove to be the best solution. It provides access to all of the institutionalized data, provides specialized tools for improving these data for transportation-planning applications, and yet makes it possible to transfer data improvements back to the

general-purpose GIS for use by others. Because of a similar architecture, the GIS-based modeling system and the institutional GIS can exchange data easily.

## *5.2. GIS-T application to modeling activities and components*

The utility of GIS-T in modeling goes greatly beyond the data preparation, visualization of inputs and outputs, and the error checking that are universally acknowledged to be a significant contribution to demand forecasting. In fact, the potential benefits of GIS-T in modeling are a great deal deeper and more fundamental, as is discussed below with respect to different aspects of the modeling process.

### *Geographically accurate road networks*

One of the important benefits of GIS-T in demand modeling is in the development of geographically accurate road networks. In current practice, GIS line layers are routinely turned into planning networks. High-quality centerline files are usually part of a municipal or regional GIS, even if transport applications were not anticipated. These can greatly simplify the process of deploying GIS-T applications. However, line layers that are prepared for other purposes or that have not been specifically prepared for transportation modeling usually require some preprocessing and special error checking. Common problems are topological disconnects stemming from undershoots and overshoots, extraneous nodes, and nodes where there are underpasses and overpasses. There may also be duplicate nodes or links, and links with the incorrect directionality. Most of these pathologies can be found with vendor-provided or user-written GIS utilities.

Line layer editing tools make it easy to add and delete links and nodes, to split links, reshape links, and to correct the topology of a network. Some facility for denoting and modifying the direction of one-way streets and bi-directional segments should also be provided. Enforcing topology in network editing would seem the best approach, although not all software works this way.

Planning networks that predate GIS usage can be corrected through a process referred to as conflation. Once considered a formidable problem, conflation tools for line layers are an important component of a GIS-T. Conflation generally refers to the merging of data from different sources so as to enhance one or both of the sources. In a GIS-T the most common conflation issue is that of substituting more accurate geography for less accurate, while retaining the attributes that are attached to the less accurate layer. When the geography changes, there may no longer be a one-to-one correspondence between the segments in both databases. The conflation problem is directly faced by anyone who has a stick representation of their network or uses a not very accurate line layer such as a TIGER file as the basis for their planning network.

An example of the results of using an interactive conflation tool is shown in Figure 7. Note that the positional accuracy of the beginning and ending nodes of the network is also improved, as well as adding the correct shape to the link. In addition to manual conflation, network-matching tools can be used to automate some of the work.

Related editing tools that are useful include doubling a line segment so that there is a separate arc in each direction, separated by a specified distance. This facilitates conversion from a centerline representation of a road to a dual representation. As shown in Figure 8, complex interchanges can be tedious to edit. Templates have been developed to perform much of the work in one step. Since most editing is done over aerial photography, final adjustments can be made to correct any variations from the template.

While this was not recognized early on by planners, significant improvements in the accuracy of networks result from this process. First, it becomes easier to identify links that have been left out of the network but that should have been included. Second, positional errors in node location are more readily seen and corrected. Third, the distances, and thus the estimated travel times, can be quite improved, leading to significant differences in model behavior and forecasts. Lastly, it is easier to correct errors in functional class. As discussed elsewhere in this volume, GPS utilization is increasingly common by transportation agencies at all levels of government. Improvements in network accuracy and in map-matching algorithms facilitate tagging GPS measurements to the correct road segments, providing a valuable source of enhanced geography and network speed information.

When displayed on a network with the correct shape, the result is more comprehensible, aesthetically pleasing, and convincing to public officials and citizens. Without a doubt, GIS technology has given transport modeling heightened credibility, even if this has not been completely warranted.

### *GIS treatment of transit networks*

An integrated modeling and GIS package offers a more detailed and accurate view of transit than that afforded by conventional planning packages. For example, bus stops are placed in their correct locations on the correct side of the street, rather than it being assumed that they are located at nodes on the street layer. Also, the routes can follow the actual street alignment. Another advantage can come from the incorporation of walking networks as part of a transit network. This has the advantage that access, egress, and transfer links need not be explicitly defined.

### *Travel choice models*

It is well known that trip-making characteristics and rates vary by location as well as with the socio-economic characteristics of individuals and households and

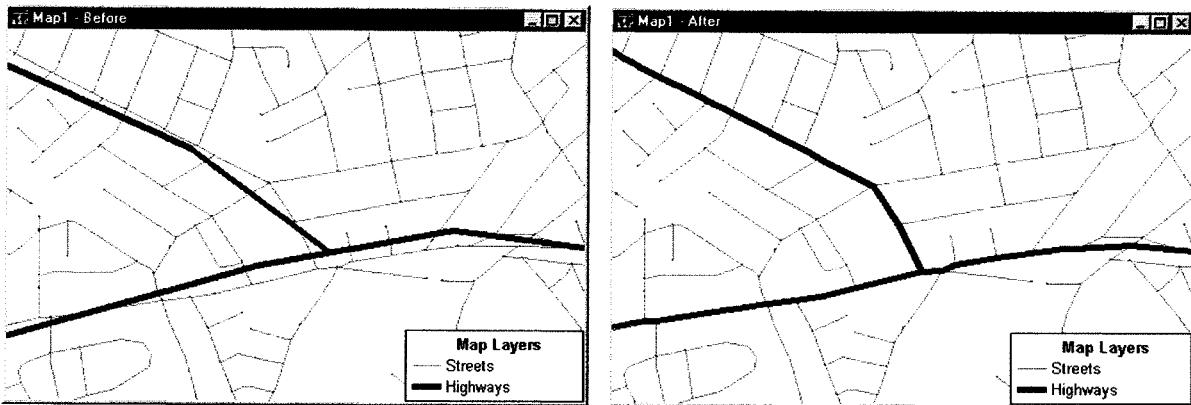


Figure 7. Network conflation.

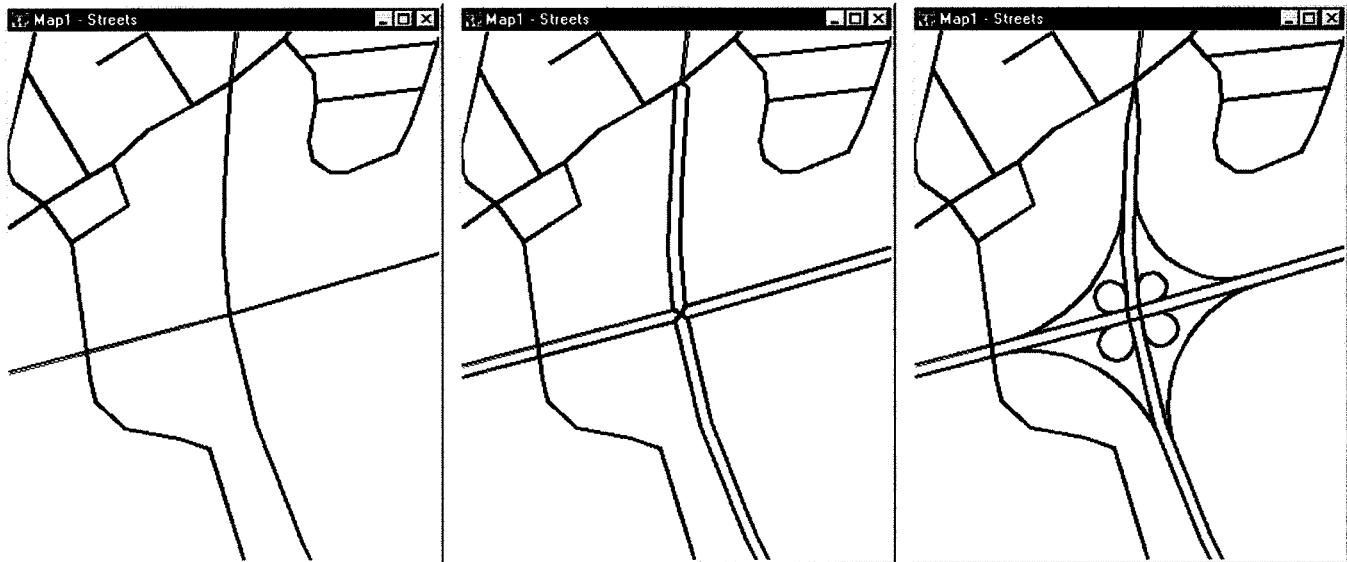


Figure 8. Line doubling and interchange editing.

transport supply. A GIS-based modeling system facilitates identifying geographic variation in trip generation and applying differential rates based upon sets of zones that combine geographic and other selection criteria. This is a better approach than using rings or other artificial characterizations of zones by location. Similarly, geographic data can be used readily in computing and incorporating accessibility measures in trip generation relationships. The GIS-T approach also permits a more refined treatment of special generators, as these can be treated as not simply being located in zones but as point locations, parcels, or larger areas that have a more intimate connection to networks. Similar logic applies to destination and mode choice in which GIS-T facilitates a more detailed characterization of the activities and activity levels found at destinations, whether located by polygonal zones, grid cells, or point addresses.

In conventional practice, choice alternatives are tagged with the network distances between origin centroids and destination centroids for each mode. Network skimming can be enhanced in a GIS framework by providing more accurate address-to-address measures of distance and travel time. For access modes, the best path can be found on the walk or drive networks and incorporated into the model.

GIS can be used to develop more precise measures of accessibility and to identify choice sets for destination and mode choice models. For example, for some trip purposes, destinations could be limited to those within an hour from the trip origin. GIS-based analysis of travel surveys can be used to identify and quantify important aspects of choice sets and trip interdependencies.

In an integrated GIS-T that supports parameter estimation, there are other possibilities for model improvement. In particular, use of spatial subsets of observations in model estimation yields differing parameters, and possibly differing significant variables. Similarly, GIS can be used to examine the spatial patterns of errors from estimated models and to compute spatial autocorrelation of residuals. Even if model estimation is to be done in a separate package, the GIS-T can be used to derive subsets of observations based upon both spatial and non-spatial criteria. Ultimately, there is the question to be addressed as to whether or not individuals who have similar characteristics behave differently in different geographic locations. If so, greater use of GIS in model estimation will become necessary.

### *GIS in traffic assignment*

GIS-T technology has ushered in a trend of using much more detailed and realistic networks in traffic assignment. As indicated previously, this results in more accurate measurements of distance and a concomitant improvement of estimated travel times. Errors in networks are more readily apparent and more easily



Figure 9. Bandwidths showing traffic flows.

corrected in a GIS-T environment. Moreover, specialized utilities can identify topological errors, duplicate links, and incorrect coding that might otherwise not be apparent from visual inspection. GIS-T techniques stimulate more detailed treatment of the type of facilities that comprise the network and facilitates the use of more tailored volume delay function parameters by link type and even for an individual link. Database information on traffic signals and signs can also be used in this process. As discussed elsewhere in this volume, GPS measurements of actual travel times can also be used in the assignment process. Collection of data on route choice using GPS can also be helpful in fine tuning assignment models.

A GIS-T can be a powerful calibration aid for traffic assignment. As shown in Figure 9, map graphics that scale the widths of network segments to reflect volumes are an efficient way to examine deviations from traffic counts and to spot links that clearly have incorrect flows. Flexible labeling of traffic flow maps help planners understand model behavior and make beneficial adjustments. Further, a database of assignment iterates can be helpful in identifying when the model is sufficiently close to equilibrium and in assuring that the maximum flow change is suitable in specific geographic areas of interest.

A GIS-T can also open the door to more elaborate means of performing traffic assignments. For example, a GIS-T can have a better representation of toll structures, particularly when these are not link based. Another example is a method for assigning trips from address to address, as has been done successfully on an experimental basis (Horner, 1998).

### *Freight modeling*

The scale independence of a GIS-T-based modeling system has facilitated national and inter-regional freight transport model applications. Among these are multimodal models that include shipments by truck, rail, air, and sea. In the USA, the Federal Highway Administration, in its Freight Analysis Framework, recently used GIS-T (e.g. TransCAD) to synthesize modal freight networks and apply a traffic model for national freight modeling for capacity analysis. Extensive use of GIS-T technology was made to integrate state level data sets, national data sets, freight data, and traffic information in a consistent network. There is also untapped potential for the application of GIS-T procedures to modeling urban truck trips. Because most urban truck trips are short and chained together on multiple pick-up and delivery tours, a GIS may be the only way to measure their trip lengths correctly.

Another application that is relatively common now is the analysis of the routing of hazardous materials. This combines shortest path calculations with buffer-derived measures of the population exposure associated with links along alternative routes. Not surprisingly, this offers a very different characterization of the spatial distribution of exposure hazards as well a chance to lessen their effect.

### *Database support for modeling*

One of the important contributions of GIS-T to transportation modeling is the provision of powerful database functionality for storing, managing, and visualizing all model data. Relational capabilities facilitate the incorporation and integration of many types of data in models, and greatly reduce the time and costs associated with data preparation. The ability to join new data tables to existing tables associated with point, line, and zone layers through a common identifier is a great productivity enhancer. Also, this makes it possible to integrate data from many organizational data systems into planning applications, and relevant data can be utilized even if they become available late in the modeling process.

Another advantage cited by Anderson and Souleyrette (1996) is the query tools that make a GIS-T superior to non-GIS-based travel models, in that data for geographic subareas of networks or regions can be easily selected and modified. The integrated geo-processing capabilities of a GIS-T for spatial queries, buffering, and polygon overlay create valuable new measures for modeling, and

spatial aggregation is trivial to perform rather than the arduous task it can be in a conventional system. Visual feedback provides a valuable additional means of identifying data gaps and errors before they are embedded in the model.

### *Support for new and advanced modeling concepts*

GIS-T technology figures prominently as a tool for implementing many advanced modeling concepts. Application of GIS to activity-based models was foreshadowed by the concept of the time-space prism, which can be directly represented in a GIS. For many years now, GIS have been considered “central to the successful development of an activity-based modeling system” (Stopher et al., 1996). The same authors note that a GIS offers a model of the region that is conceptually similar to that utilized by travelers in making activity and travel choices.

Perhaps the first step in providing GIS-T support for activity modeling is processing of diary data into a form that is amenable for analysis. One way to do this entails building routes that correspond to the distinct tours that individuals make. The stops on the tours are the locations of the stationary activities that are pursued. Stop attributes can capture the nature of the activity and its duration. The route segments correspond to the trips that are made, and can similarly be characterized by attributes such as departure time, arrival time, trip purpose, mode, and whether the travel is unaccompanied or not. Route attributes can include the number of stops, the departure time, the primary travel purpose, primary mode, or all modes utilized. Tours as database entities can be analyzed through tabulation and statistical measures or utilized in the estimation of choice models. Visual analysis should also be accorded an important role in this form of model building, and thus the ability to visualize travel geography, as shown in Figure 10, is a major advantage of the GIS-T approach.

The output of activity models and tour-based models can similarly be processed, displayed, and assigned to the appropriate multi-modal travel networks. Much of the same technology can be harnessed for the purpose of collecting travel survey data utilizing Internet-based tools that record the actual routes utilized.

All of the components of the microsimulation approach to modeling travel can be implemented in a GIS-T, although some extensions may be required. By representing the locations of residences, workplaces, and other potential trip destinations as points or small areas, these models can achieve a freedom from the use of zoning systems and the aggregation (and aggregation bias) typically associated with zonal models.

Population synthesis, in which hypothetical households are derived consistent with some zonal aggregate marginal distributions, can be performed in a GIS environment, and each household can be located in a specific network location.

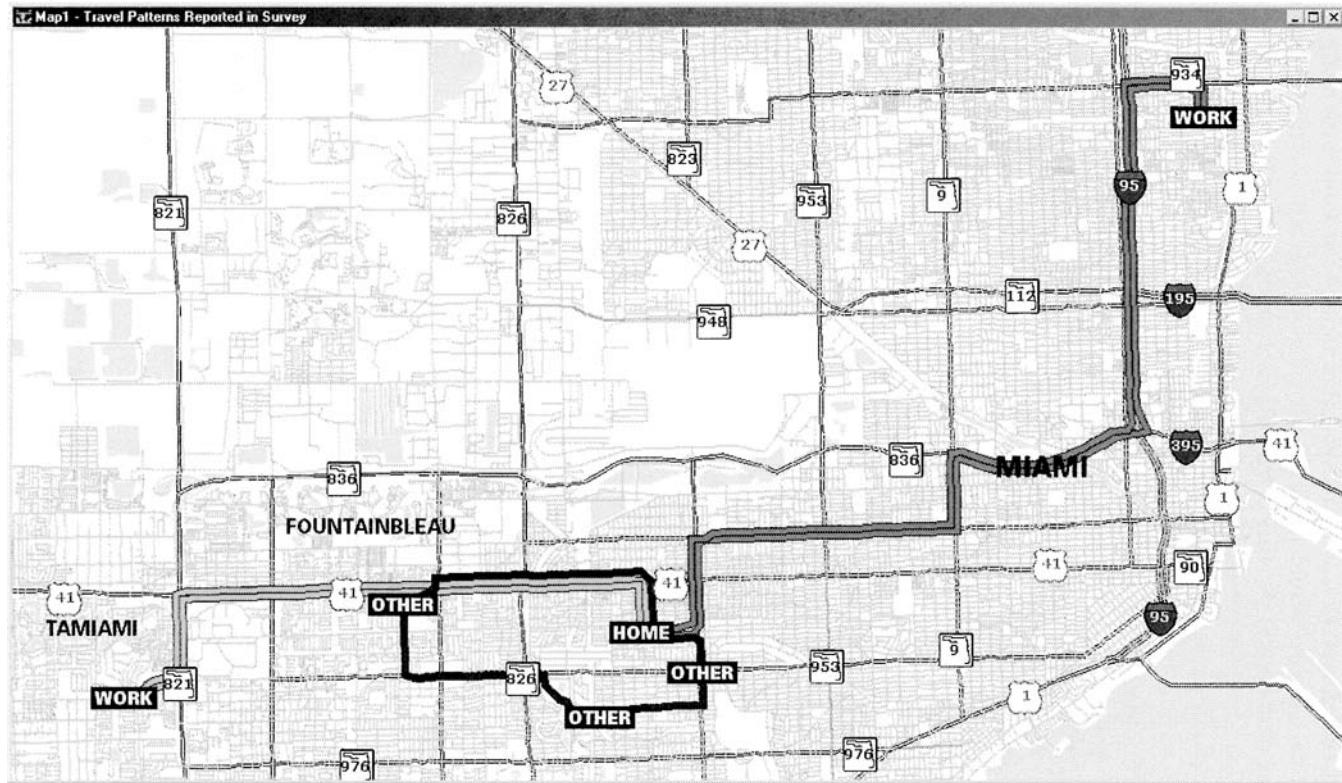


Figure 10. Travel geography viewer.

Models of trip frequency, destination, and mode choice can be applied to the database of synthesized individuals, and the outputs can be transformed in the manner described above. Furthermore, the advent of GIS-T-based traffic simulators will enable carrying the microsimulation all the way through from trip generation to network assignment.

A GIS-T has been rarely applied to the types of traffic models favored by traffic engineers. Typically, the analytical tools of the traffic engineers have relied on schematic graphic representations of networks rather than geographically accurate representations. Of course, there have been linkages between a GIS and software for signal timing and network simulation. However, these have been indirect, one-way, and incomplete in that the GIS-T provided little or no support for these applications; it was simply used to display some aggregate results from the traffic software. This is changing as GIS-T technology is further extended to handle lanes and signals and other objects associated with more accurate characterization of transportation systems and the specific entities found in advanced traffic models. Under development for 5 years, the TransModeler software (Yang and Slavin, 2002) fully integrates GIS and traffic simulation with an extended GIS data model that incorporates both the time dimension and vehicle dynamics.

## 6. Concluding remarks

GIS and GIS-T are clearly now well-established tools of planners and are being deployed to support the modeling endeavors of greatest importance to transport and land use planners. Whether linked or integrated with models, GIS technology brings many valuable new and enhanced functions that can be used in modeling, data preparation, and visualization. In addition to geo-spatial processing and selection, the powerful database capabilities provide a better software platform than software with more limited file management and graphics tools. Some of the modern conveniences of use of a GIS in planning and modeling stem from the relatively higher standard set and functionality achieved by GIS software, which has a wider audience and must conform to a higher standard in implementation and documentation than niche software such as that for land use or transport models. Integrating GIS and models brings each to a higher level, and having all the GIS and modeling tools in one place maximizes their effective and creative use.

Both GIS-T and planning models are evolving to address many of their prior limitations. Foremost among these are improvements necessary to model temporal dynamics. This includes the data and model management required to describe and to forecast minute-to-minute changes in travel and traffic and year-to-year changes in land use. Rapid expansion of available computing power will

make greater detail and complexity feasible, but will not relieve us of the need to deepen our understanding of the relationships between transport and land use and to put greater geographic intelligence in the planning process. Also, every modeling approach has its own software requirements, and new models will continue to expand the requirements for integration with GIS. Fuller development of specific GIS-T applications will also be necessary to bring the utility of the technology to planners and engineers. While GIS-T technology will not make planning or modeling easy, it will make them better.

## References

- Anderson, M.D. and Souleyrette, R.R. (1996) "A geographic information system based transportation forecast model for use in small urbanized areas," *75th TRB Annual Meeting*, Paper. Washington, DC.
- Caliper Corporation (2002a) *Travel demand modeling with TransCAD 4.5*. Newton: Caliper Corporation.
- Caliper Corporation (2002b) *Step 2 for Clark County: household microsimulation for transportation policy analysis*, Report. Las Vegas: Southern Nevada Regional Planning Coalition.
- Dueker, K.J. and T. Ton (2000) "Geographical information systems for transport," in: D.A. Hensher and K.J. Button, eds, *Handbook of transport modelling*. Oxford: Pergamon.
- Green, M. and R. Flowerdew (1996) "New evidence on the modifiable areal unit problem," in: P. Longley and M. Batty, eds, *Spatial analysis: modeling in a GIS environment*. New York: Wiley.
- Harvey, G. and E. Deakin (1996) "Description of the STEP analysis package." Unpublished.
- Horner, M. (1998) *Direct household assignment: an exploratory study of simplifying urban travel forecasts using GIS*. Charlotte: Department of Geography, University of North Carolina.
- Landis, J. (2001) "CUF, CUFII, and CURBA: a family of spatially explicit urban growth and land use policy simulation models," in: R. Brail and R. Klosterman, eds, *Planning support systems*. Redlands: ESRI Press.
- Lee, D.B. (1973) "Requiem for large-scale models," *Journal of the American Institute of Planners*, 39:163-178
- Stopher, P., D. Hartgen and Y. Li (1996) "SMART: simulation model for activities, resources, and travel," *Transportation*, 23:293-312.
- Waddell, P., T. Moore and S. Edwards (1998) "Exploiting parcel-level GIS for land use modeling," in: *1998 ASCE Conference on Transportation, Land Use, and Air Quality: Making the Connection*. Portland.
- Waters, N. (1999) "Transportation GIS: GIS-T," in: P. Longley, M. Goodchild, D. Maguire and D. Rhind, eds, *Geographical information systems: principals, techniques, applications, and management*, 2nd edn. New York: Wiley.
- Wegener, M. (1995) "Current and future land use models," in: *TMIP Land Use Modeling Conference Proceedings*. Washington, DC: US Federal Highway Administration.
- Yang, Q. and H. Slavin (2002) *High fidelity, wide area traffic simulation model*. Newton: Caliper Corporation.

*Chapter 20*

## THE ROLE OF GIS IN ROUTING AND LOGISTICS

JOHN C. SUTTON

*Cambridge Systematics, Chevy Chase, MA*

JOHAN VISSER

*Institute for Housing, Urban and Mobility Studies, Delft*

### 1. Introduction: why use GIS in routing and logistics?

According to the Council of Logistics Management (CLM), logistics is the process of planning, implementing, and controlling the efficient, effective flow and storage of goods, services, and related information from the point of origin to the point of consumption for the purpose of conforming to customer requirements. To accomplish this requires the optimal location and number of depots and warehouses, as well as the efficient management of vehicle fleets to meet delivery schedules together with effective planning of vehicle routes. Today, logistics and routing is an integrated activity involving multiple modes and trans-shipment at intermodal facilities. The use of the latest routing and logistics tools that can help plan, allocate, and track vehicles and shipments in space and time is necessary to meet customer requirements for supply chain management and just-in-time delivery.

The logistics industry has responded to this challenge, leading to the widespread use of information and communications technology. With the growing use of Global Positioning System (GPS) and automated vehicle location (AVL) technologies, the attraction of a computer-based mapping technology such as geographic information systems (GIS) has increased. GIS integrate spatial and attribute data to create a powerful database and mapping system; they are designed to work with location-based systems, and are therefore an attractive choice for logistics providers. The use of routing software is still not common practice for the majority of companies in the transport industry, and most routing and logistics software is not based on GIS. Routing and scheduling is a complex problem, especially when managing multiple vehicles, organizing complex tours, trying to accommodate intermediate stops on-the-fly, or routing with backlogs (Bodin et al., 1983). While GIS may not be able to solve all these problems, their use in logistics and routing is growing. The widespread availability of spatial

data, including digital networks, combined with the visualization and mapping capabilities of GIS add to their attractiveness, as described below.

There are several steps involved in transportation logistics – from route planning to delivery – as depicted in Figure 1. GIS can help at each stage in the process, beginning with the geocoding of customer locations ( $x$ ,  $y$  coordinates) using the geographic database of streets. Together with the locations of the shipment origin (e.g. a freight warehouse), these are used in building the origin-destination trip matrix. The street information stored in a GIS can be used to calculate the distances between all customers as well as the network travel times, providing the basic information to the assignment and route improvement steps. Using distance and other customer information, such as imposed time windows and special delivery requirements, the assignment and route improvement steps generate the final routes for delivery. This sequence of steps applies to most logistics business whether transporting freight or passengers.

By combining the customer attributes with the spatial data, a GIS provides a powerful database and mapping system that can display the final routes on a map as well as other information of use to the delivery drivers or the operations managers. A GIS thus allows the effects of routing and logistics decisions to be seen visually, projected on to a map. Using a map to assist in the route choice and scheduling selection process improves the efficiency and effectiveness of the routing and logistics planning process. An example of a GIS-based logistics mapping application is depicted in Figure 2; the route in the figure can be matched with, for example, the truck time-windows of operation to determine the shortest route, or any scope to add additional stops to the route determined. This ability to query and display transportation spatial and attribute data in multiple formats is one of the major advantages of utilizing GIS.

## **2. GIS routing and logistics capabilities**

In order for a GIS to perform routing and logistics procedures, it needs some basic building blocks:

- A GIS is built upon the foundation of a digital base map that includes roads, railways, rivers, and other transportation features. The linear features are connected together in a series of links and nodes (edges and junctions) that comprise the basic network for routing. The links or arcs can be coded as one- or two-way features, depending upon the level of detail needed for navigation and mapping purposes.
- In addition to the navigation network a navigable database is required to store the attributes that describe the network features, including one-way streets, impedance values, such as speed limits, and turning restrictions at junctions.

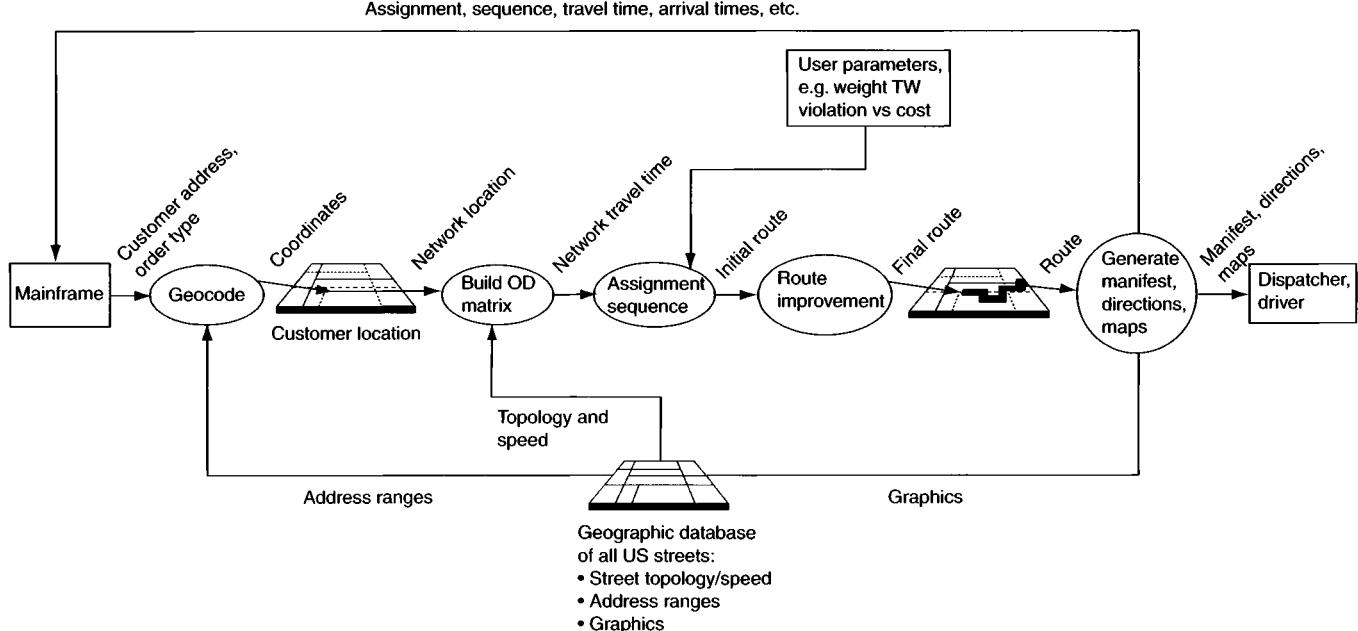


Figure 1. Applying GIS to the logistics planning process. (After Wiegel and Cao, 1999; reproduced by courtesy of ESRI.)

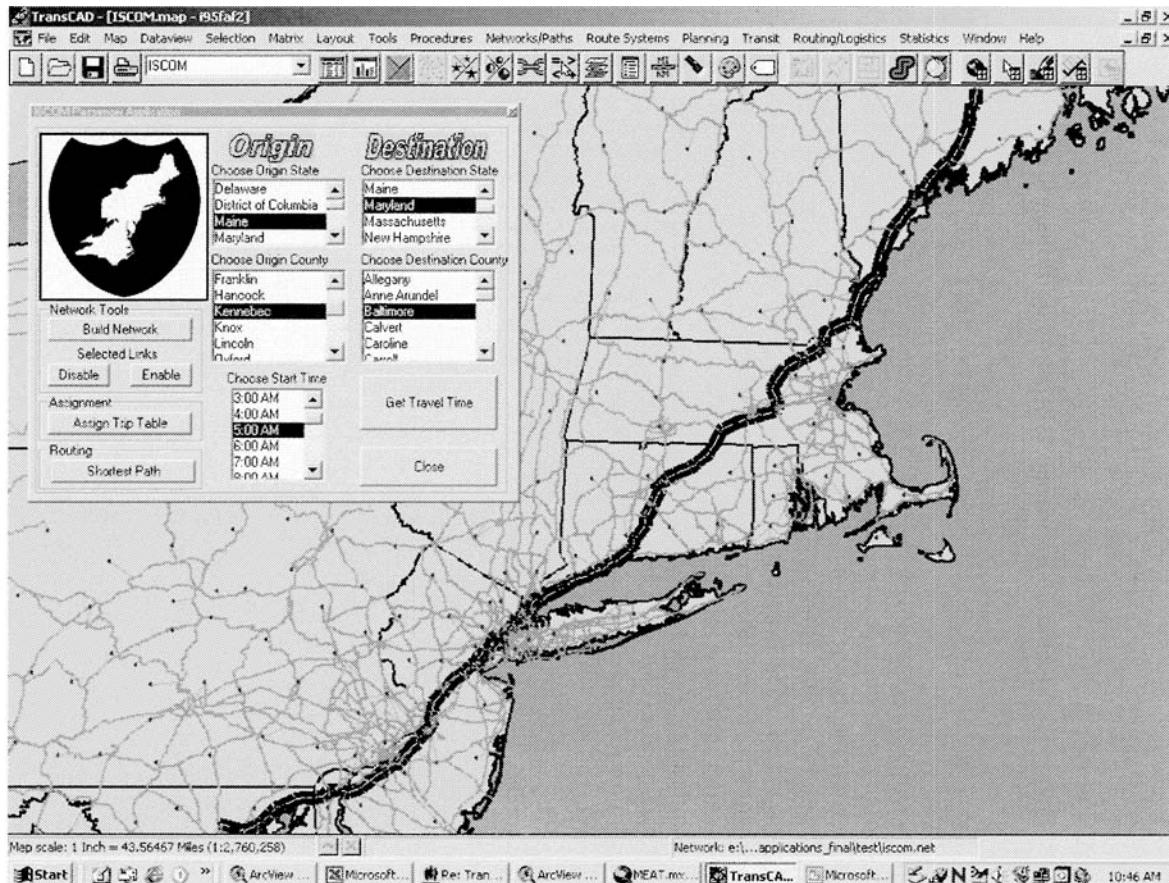


Figure 2. GIS routing interface.

- A database is also used to store the origins and destinations flows that are compiled in a trip table for assignment to the network.

Once these building blocks are in place, the GIS can be used to perform a wide variety of routing and logistics functions.

### *2.1. Vehicle routing/dispatching*

A GIS provides a rich set of tools that solve various types of pick-up and delivery routing problems (Caliper Corporation, 1996). These tools are used to prepare input data, solve the routing problem, and provide tabular and graphical output of the resulting routes and vehicle schedules. GIS-based procedures can solve many variations on the classic vehicle-routing problem, including restrictions on the time when stops can be made, the dispatching of vehicles from multiple depots, and the use of vehicle fleets comprising multiple vehicle sizes. Some of the more advanced GIS programs are also capable of solving problems involving mixed pick-up and delivery. In addition, users can edit the routes interactively via the map interface to add or remove stops and then perform a re-optimization of the route so as to minimize time-window violations.

### *2.2. Arc routing*

Arc routing problems are a class of problems that involve finding optimal paths through a set of links in a transportation network. Arc routing has a large number of public and private sector applications, including garbage collection, school bus routing, mail delivery, and other door-to-door operations. In a typical arc routing problem, people or vehicles are dispatched from one or more facilities to traverse a set of service links. The solution to an arc routing problem is a set of one or more routes that cover all the service links with the minimal amount of downtime.

### *2.3. Network flow and distribution analysis.*

A GIS can be used to solve network flow problems, which require the efficient delivery of goods or services:

- The transportation problem involves identifying the most efficient way to service a set of destinations from a set of origins. For example, a company may be interested in finding the least-cost solution for shipping commodities from its warehouses to its retail locations.

- The minimum cost flow problem is a more general version of the transportation problem that takes link capacities into account. Examples include bridge height or weight restrictions, or congested links that slow travel time. The procedure can be used to find alternative paths when capacity constraints make it impossible to utilize the shortest path between an origin and a destination.
- Matching problems try to find the best one-to-one matching between two groups of objects where there is some quantitative measure to be minimized or maximized. An example is how to assign work to service depots in the most efficient way. Figure 3 illustrates an example of how a GIS can solve this matching problem.

#### *2.4. Location and allocation models*

GIS procedures for regional partitioning, clustering, and facility location have broad applications in transportation and marketing. Clustering routines assemble customers, facilities, or areas into groups that are compact and can be serviced efficiently. Districting models group census tracts, ZIP/postal codes, counties, or other regions into territories that are compact and balanced. Location models evaluate the costs and benefits of any number of proposed facility locations.

##### *Service area definition*

A GIS provides powerful automated procedures for defining territories:

- Districting or partitioning involves creating groups of features in a layer based on proximity or measures of similarity. The partitioning procedures in a GIS support applications in service territory alignment, marketing allocation, political redistricting, and many other disciplines. The partitioning model attempts to produce districts that are contiguous and evenly balanced.
- Clustering is the grouping of features into compact clusters where there may also be limits on the size of each cluster or capacity. The clustering procedure can be used to solve problems in many disciplines such as vehicle fleet management or the allocation of students to school buses.

##### *Site location analysis*

Site location problems involve choosing the best location for one or more facilities from a set of possible locations. A GIS can address many types of location problems, including determining the number of facilities that are required to guarantee a prescribed level of service, for example the number and location of fire stations; assessing the trade-offs in the cost of adding a facility with the

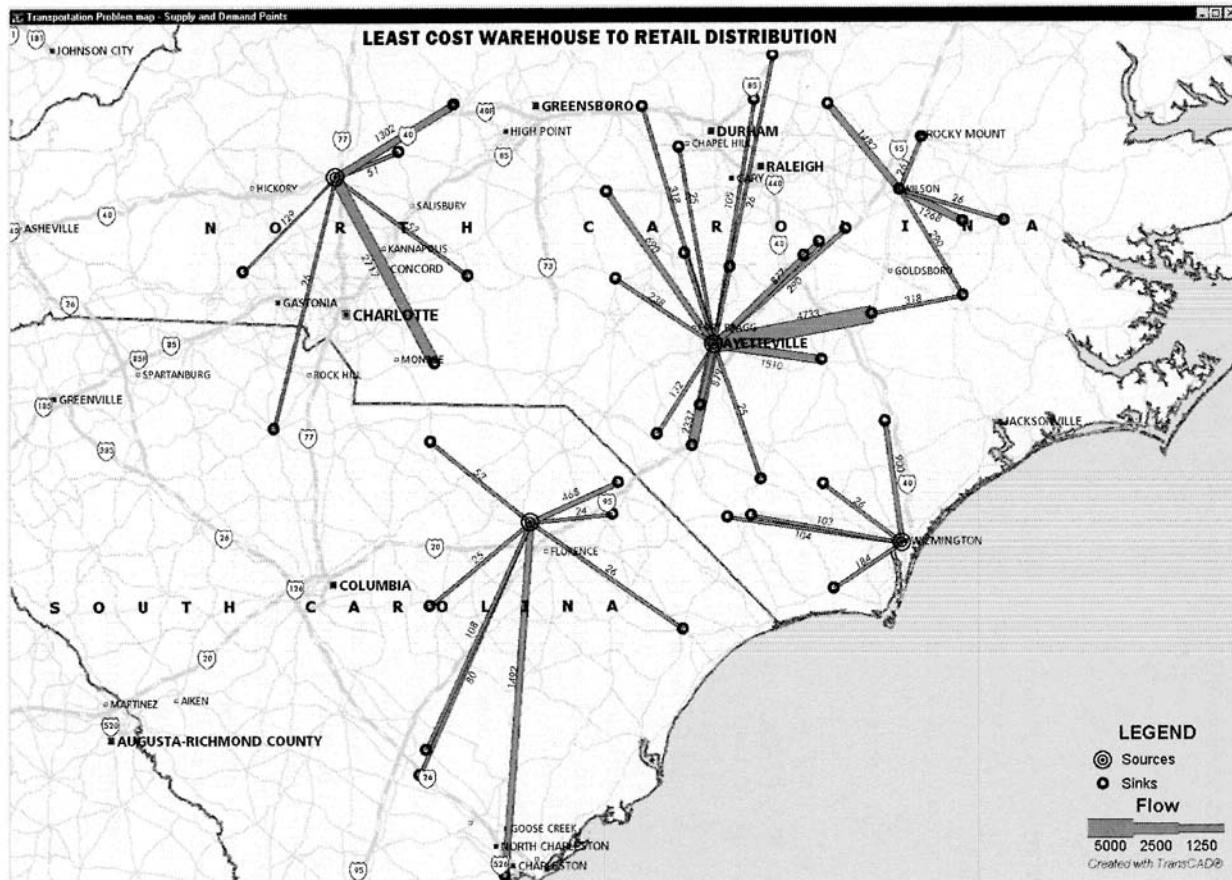


Figure 3. Example of service areas defined by depot locations. (Reproduced by courtesy of Caliper.)

potential revenue benefit; establishing thresholds for facility location such as industrial sites or landfills that need to be located relatively far from population centers; and evaluating the competition from existing facilities.

### **3. Logistics issues**

The routing and logistics capabilities of GIS are widely recognized in the logistics industry. GIS are used at the levels of operational, tactical and strategic decision-making in transportation planning. Operational planning means the daily trip or route planning of a transport company from one or multiple depots. When transporting goods from one location to another in a cost efficient way, and at reasonable standards, the operator has to determine how many vehicles are required for the deliveries as well as develop a schedule for each vehicle. The vehicle routing/dispatching tools of a GIS can be used to optimize and assign routes to vehicles (vans, trucks), based on variable sets of destinations. These procedures develop routes for making pickups or deliveries using vehicle dispatching from a central location. Decisions for route schemes with a fixed large number of destinations can be based on the arc routing tools of a GIS, as described above.

In the case of a multi-depot situation, receivers are usually assigned on a yearly basis to a specific depot. The network flow tools of a GIS can be used to match the set of receivers, such as retail shops or other businesses, with a set of warehouses. This way, decisions on closing or opening new depots or shifting trucks from one depot to the other can be supported by detailed information. In order to define service areas for depots, partitioning and clustering tools, as described above, can be used.

At the strategic level, optimizing the number and location of warehouses or trans-shipment terminals is a very important issue in logistics. Decisions on these facilities are normally made for the long term, so making the best decision on location can be critical to the success of the logistics solution. For these reasons, site location models, also called facility location models, are sometimes employed to support the decision-making process. These models handle multiple objectives, including improving the level of service, reducing the cost of service, and profit maximization. Facility location models can typically be classified (Taniguchi et al., 2001) into three types:

- continuous location models;
- network location models;
- discrete location models.

Discrete location models are the most practical, because it is assumed that the number of candidate locations is finite (as depicted in the example in Figure 3).

Network location models can be used to optimize service patterns, based upon network length or travel time criteria. An example might be determining bus routes through an urban area that maximize passenger accessibility while at the same time minimize in-vehicle travel time. In this example, the walk time costs of passengers have to be balanced against the vehicle cost of the bus operations.

Continuous location models are more flexible but also more difficult to solve. This occurs in emergency response, where police vehicles or ambulances are not fixed in their location; or a demand-responsive taxi service, which is dispatching taxis to meet short-term requests. The analysis of these patterns of demand can indicate the level of service that needs to be provided in each service area to meet average demands.

#### 4. Public policy-making

Related to the use of GIS in logistics, is the use of GIS as decision support systems in public policy-making, such as freight transport. Planning and regulating freight traffic is an important public policy issue. The potential of GIS here is not so much assisting the transport operator but rather supporting the public authorities in their modeling of, and planning for, urban freight movements, including making infrastructure investment decisions.

In the policy-making process, different studies can be carried out utilizing GIS. One example of this is freight traffic analysis using GIS. Figure 4 shows an example of integrating truck count data that can be visualized with GIS in a number of ways: by volume, by percentage of total traffic and by percentage growth. The map depicts the growth rates of truck traffic on major roads in the Fort Lauderdale area of Broward County, Florida. This type of data can be used to evaluate truck impacts and design mitigation measures, such as designating truck routes, and where to impose truck restrictions. Collecting and analyzing this data over a number of years enables trends to be determined. This is useful for calibrating travel demand models for predicting future truck traffic volumes.

Likewise, GIS are an especially important tool when environmental impacts (noise, air pollution or energy consumption) have to be estimated. These data have different spatial extents, so overlaying them with traffic volumes in GIS is an excellent method.

Another important issue in public planning is to find patterns in the location of logistic centers (distribution centers, consolidation points, warehouses, terminals, and so on) for an optimal planning of public infrastructure for freight transport. This includes the location of access roads, rail terminals, or waterborne terminals.

These GIS-based routing systems can also be used to evaluate the impacts of major incidents that may close down a section of highway. This is of critical concern, following the terrorist attacks of September 11, 2001. A number of

studies employing GIS have been undertaken to assess the logistics problems in evacuation and incident management. The ability of GIS to integrate a variety of disparate data in a single database and mapping system make them an important tool in decision-making.

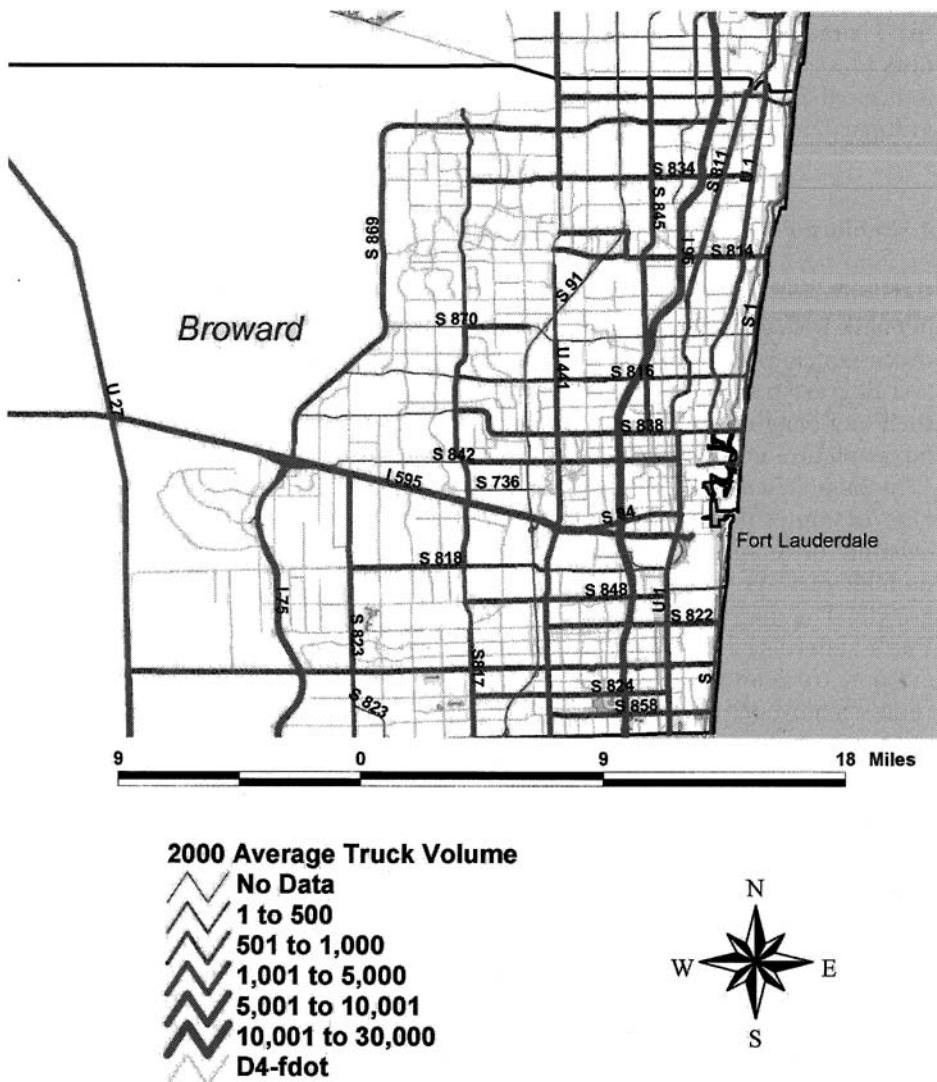


Figure 4. Truck count mapping system in Florida.

#### 4.1. Freight modeling

It is common practice in transportation planning to make use of transport modeling and traffic simulation software tools. GIS software can be linked to these models in order to display the model outputs and, in some cases, perform network building and editing for input to the transportation model. The TransCAD GIS, for instance, provides tools for transportation modeling and traffic simulation. Although the main focus in transport models is on the modeling of flows of passengers, they can be designed to model freight transport flows. One example is the city of Sienna (Italy). Here, a feasibility study was carried out to improve the efficiency of goods distribution in this historic city (Valentini et al., 2001). A transport model was integrated with environmental models in a GIS. With information on destinations, freight traffic data, tourist flows (Sienna is a popular tourist destination), and estimation of freight volumes, the most optimal distribution patterns and routes were determined. GIS can thus be used to integrate different types of data and models in a single routing and logistics application.

Another example of using a GIS to model freight transport flows can be found in Delft, the Netherlands (Visser and Maat, 1996), as depicted in Figure 5. In Delft, as in other cities, the routing of freight is not only a matter of finding the shortest or fastest routes through urban areas – vulnerable areas, such as residential neighborhoods, hospitals, and so on, should be avoided. Traditional vehicle routing methods are inefficient, but with GIS the aforementioned factors can be included in the procedure to find the optimal routes.

In the USA, GIS are used in several states to route hazardous materials. In most US states, truck permits are issued for overweight and oversize vehicles that may be carrying exceptional or dangerous loads. In these cases, the vehicles have to be routed to avoid low overpasses or bridges with weight restrictions. A database of these constraints can be used with the GIS to determine routes that avoid these features.

#### 4.2. Spatial studies

Button et al. (2001) studied the spatial distribution of freight transport logistics centers in large urban areas (Detroit and Washington–Baltimore). The extent to which the spatial distribution of urban freight centers differs between various urban forms was investigated with a GIS. The research method in this study employs a technique known as “near-neighbor analysis,” which is a popular geostatistical method for determining the spatial correlation of features. It seeks to identify situations where a distribution of points, i.e. locations, is clustered, random, or scattered. The study highlights the location and spatial distribution of



Figure 5. Freight traffic modeled in the city of Delft (Visser and Maat, 1996).

the trucking/courier and warehousing industry in each of the urban areas. This example illustrates how GIS can be used to research the geography of freight distribution centers in urban areas and determine the patterns associated with different urban forms, which can then be applied to other cities at different stages of their development.

The changing geography of urban areas, especially urban sprawl, together with increasing levels of traffic congestion, present a number of challenges to city planners, shippers and retailers. City planners use GIS to model land use and transportation interactions in order to more fully understand the planning implications of changes in the spatial distribution of activity centers. The goal is to develop policies and plans that will create a better balanced land use and transportation system. In some cases, this may be in conflict with shippers and retailers who manage the supply chain. These are difficult issues with several dimensions, including the correct balance to achieve between public goals and consumer choice. GIS alone cannot solve these problems, but they can help to inform the debate by compiling data and performing spatial analyses that illustrate the consequences of planning decisions, whether in the public or private sectors.

## 5. Real-time routing and logistics

### 5.1. From static to dynamic information

A combination of new technologies, including GPS and mobile or wireless communications, make it possible to use real-time information within GIS. Wireless data communication networks can be applied to link vehicles to the base server, while GPS makes it easier to determine the location of vehicles.

Real-time data can be used for dynamic routing and scheduling. A range of dynamic information can potentially be incorporated, such as actual traffic information, roadworks and accidents, actual travel speeds or travel time, weather conditions, and customer information. The Internet can be used to distribute information to shippers and dispatchers from the GIS servers. GIS are now integrated with Internet technologies through internet map servers, and these can be linked to mobile devices including personal data assistants (PDAs) that can be used to communicate with the central data server. Also, via web-based technology, real-time data can be integrated in databases related to road, traffic, and environmental conditions, and these can be validated independently, and accessed independently or as part of a comprehensive traffic management system (Taniguchi et al., 2001). Advanced logistic systems may therefore be introduced, making freight transport more efficient and improving customer services.

### *5.2. Convergence of GIS and location aware technologies*

GIS software has matured to the point where it can now be integrated with other information and communications technologies to support the latest location aware technologies (LATs), being supplied by location-based service (LBS) providers. New LBS companies are entering the market for real-time traffic data and vehicle tracking services. These companies are embedding basic GIS applets into their products, with simple mapping and query tools. Map vendors in Europe, Asia, and North America are providing value-added services to their digital products to offer routing and location "Geo-Yellow Pages" services. Some of these are subscription services, such as General Motors On-Star in-vehicle telematics service. This service appeals to individual subscribers or small distribution companies, who have less need for an advanced GIS-based routing and logistics program. These pay-as-you-use wireless accessible services are forecast to grow as telecommunications technologies expand their coverage and penetrate ever more business applications.

Generally speaking, the new LATs are proprietary systems that embed some GIS components but do not support GIS queries or analysis beyond the specific domain of the application. However, they can export their data in formats that can be read by other GIS packages. There are a number of international efforts underway to standardize transportation data exchange and communications protocols through the International Organization for Standardization and others. Many of these are driven by the needs of the intelligent transportation systems (ITS) vendors, who see a global market for their products and services. The logistics industry is already embracing ITS technologies in real-time routing and dispatch, vehicle tracking, and suchlike, and this is likely to be an important influence on the development of the next generation of GIS software for the transportation industry.

## **6. Software**

At this time in the routing and logistics industry, dedicated software is more popular than GIS software. Vehicle-routing software entered the commercial market much faster than GIS software and is easier to customize to customers' needs. Although dedicated software has many of the features of a GIS, the set of available tools is smaller and far more specialized. Not all GIS programs have tools for vehicle routing, arc routing, network flow analysis, or location/allocation models, although in many cases these can be programmed.

There are a number of ways that GIS can be applied in vehicle routing and scheduling. First, a GIS can be used as an interface to the routing and scheduling procedures. In this case, the logistics programs run the procedures, with the GIS

being used to manage the network data and display the results. This can be achieved using a dynamic data exchange (DDE) protocol and establishing an application program interface (API) link between the two packages. Examples of this type of integration are the Route Smart Software with ARC/INFO GIS, and OPCOM's vehicle-routing utilities with MapInfo (Taniguchi et al., 1999). Commercial map vendors, such as Navigation Technologies, Tele Atlas, and Rand McNally, also provide their own proprietary routing extensions that work with their digital networks, otherwise they can export their data in a GIS format. The map vendors also supply their data to Internet service providers. Internet service providers, such as MapQuest and Lycos, provide free address matching and routing between origin and destination pairs. These simple routing and mapping applets are useful for single trip planning but not powerful enough yet for logistics or service planning of multiple vehicles or multiple routes.

More advanced GIS software has routing and logistics functions built-in or provided through extensions to the core program. Routing and scheduling can be performed within the GIS. Many desktop GIS have a set of routing procedures in-built or allow user defined modules.

There are several specialist GIS products available for logistics and routing, including:

- ArcLogistics Route (from ESRI), which provides delivery routes within time-windows and optimizes multiple vehicles and delivery schedules ([www.esri.com](http://www.esri.com));
- Network Analyst – an extension to ESRI's ArcGIS product that includes shortest path, tour, and catchment area functions;
- TransCAD GIS (from Caliper), which provides a comprehensive set of routing and logistics tools as part of its core functionality, including arc routing, vehicle routing and dispatching, network flow and distribution analyses, and location-allocation models ([www.caliper.com](http://www.caliper.com));
- Geomedia Transportation Manager and Transportation Analyst (from Intergraph), which provide advanced network analysis functions that can be customized for routing and logistics applications ([www.intergraph.com](http://www.intergraph.com));
- MapInfo, which provides a routing J server platform that can be embedded in a web-based or Microsoft Windows-based application ([www.mapinfo.com](http://www.mapinfo.com)).

This list is not exhaustive, but is representative of the products that are available from GIS vendors.

### *6.1. GIS limitations*

While GIS have proved to be an excellent tool for routing and logistics, there are some limitations:

- Networks can be built as geometric representations of the base map or as schematic representations tied to the underlying base map. Examples of the latter include airline or ocean shipping networks where routes are not fixed by linear features. Sometimes, schematic networks (“stick diagrams”) are used for routing on surface networks where performance may be affected by the complexity of the geometric network, such as when routing over large networks. In these cases, a simplified schematic or logical network may be preferred for the routing algorithm execution, but with the results displayed on the complete base-map geometric network.
- When networks are built in a GIS, all the spatial features are topologically connected and managed within the GIS. This “spatial intelligence” is one of the key features of GIS and of great benefit compared with other mapping software, which simply display “dumb graphics.” However, a disadvantage is that transportation features may be connected in violation of real-world conditions, such as where an overpass is connected to the underlying road. The user needs to be aware of these issues when coding the network features, or illogical and erroneous routing paths will be created.
- Logistics takes into account the temporal dimension as well as spatial dimension, to optimize vehicle schedules and deliveries within the spatial and temporal constraints. GIS software to date has not provided robust functions for optimizing with time constraints, although as described in the introduction, GIS approaches that include time scheduling are beginning to be developed. Most non-GIS supply chain management and routing software packages already deal with scheduling problems, and these are often more advanced implementations than spatial routing or location-allocation problems.
- The most complex routing and logistics problems are not addressed in GIS. Because GIS manage data spatially, the routing constraints on the network or at junctions may be more problematic to specify than in non-GIS programs. These generally have more flexibility in coding network solutions and can therefore manage more complex situations.
- The performance of GIS programs may be slower than non-GIS programs because the latter do not carry the spatial topology overhead of GIS software. This is related to the first limitation listed above, and is one of the reasons why some programs perform routing on more simple networks.

Generally speaking, the lower-cost desktop GIS software packages provide only basic network analysis capabilities. The full range of routing and logistics capabilities are found in the more advanced GIS programs. As with other products, users get what they pay for.

## 7. Conclusion

Despite the limitations, GIS software has evolved to provide a rich set of network and spatial analysis tools for routing and logistics. The range of applications that GIS can address is impressive and increasing with each advance in the technology. The early versions of GIS software were proprietary programs that had their own scripting languages with commands that took a long time to learn, and required specialist skills to program. Since then, GIS have moved into the mainstream and become more “open” with respect to configuration with operating systems, especially Microsoft Windows, and GIS data transfer standards. These trends bode well for the GIS industry, and have extended the reach of GIS into most areas of the transportation business. The logistics industry mirrors these changes, and has already benefitted from the advantages that GIS bring:

- The ability of GIS to integrate spatial data with other data is a powerful feature. Examples include the building of navigable networks and databases for routing and logistics applications. Further, GIS can be integrated with off-the-shelf database management systems, avoiding the need to store data in proprietary formats. This flexibility means that network attributes can be changed by users according to real-world conditions on the network. In stand-alone routing software, the network data may be frozen, and this is a serious limitation when errors need to be corrected. GIS still require some knowledge and training in how to set-up and manage spatial data, but the latest GIS technology is more straightforward to implement, and has a shorter learning curve.
- The cost of GIS software and associated training has decreased in real terms. This is especially the case with desktop GIS programs, which provide a lot of functionality at a reasonable cost. More advanced routing and logistics functions are available at additional cost through program extensions, or more advanced versions, but the prices are still competitive compared with proprietary non-GIS programs. The growth of GIS in logistics has spawned a number of third-party vendors and consultants that provide technical support and custom applications. Thus, the technology and the support services have gained maturity.
- GIS have excellent technical capabilities that allow complex routing, scheduling, and logistics functions to be performed. Among the more advanced GIS programs, the range and flexibility of the logistics tools cover most functional areas. In some cases, the functions are at the cutting edge of theory, demonstrating uses that extend the role of logistics into many organizations that previously would not have considered such tools. In short, GIS are making popular what were once considered specialist applications.

- As these benefits converge, the critical mass of available spatial data and easy-to-use processing tools is bringing the cost of logistics applications within the realm of even small organizations. GIS technology, therefore, is a catalyst in growing the routing and logistics market place, and can claim some credit for advancing the state of the art in this area. The integration of GIS with GPS and other location-based technologies is an exciting prospect and one that will benefit further the application of GIS in routing and logistics.

## References

- Bodin, L., B. Golden, A. Assad and M. Ball (1983) "Routing and scheduling of vehicles and crews: the state of the art," *Computers and Operations Research*, 10:63–211.
- Button, K., R. Kulkarni and R. Stough (2001) "Clustering of transport logistics centres in urban areas," in: E. Taniguchi and R.G. Thompson, eds, *City logistics II*. Kyoto: Institute of Systems Science Research.
- Caliper Corporation (1996) *Routing and logistics with TransCad 3.0*. Newton: Caliper Corporation.
- Taniguchi, E., R.G. Thompson and T. Yamada (2001) "Recent advances in modelling: city logistics," in: E. Taniguchi and R.G. Thompson, eds, *City logistics II*. Kyoto: Institute of Systems Science Research.
- Valentini, M.P., P. Lacquaniti and G. Valenti (2001) "Methodology and results of a study on logistic schemes in Sienna," in: E. Taniguchi and R.G. Thompson, eds, *City logistics II*. Kyoto: Institute of Systems Science Research.
- Visser, J.G. and C. Maat (1996) "A simulation model for urban freight transport with GIS," in: *24th European Transport Forum*. London.
- Wiegel, D. and Cao, B. (1999) "Applying GIS and OR techniques to solve Sears technician-dispatching and home-delivery problems," *Interfaces*, 29:112–130.

*Chapter 21*

## GIS AND THE COLLECTION OF TRAVEL SURVEY DATA

STEPHEN GREAVES

*Monash University, Clayton*

### 1. Introduction

Travel data are integral to transportation planning and policy decisions. Typically, these data are collected through some type of participant-based survey, which is a notoriously expensive and complex planning activity (Stopher, 2000). A fundamental requirement for useful travel survey data is that they are accurately related to specific geographic locations. This is achieved through a process known as geocoding, which involves the assignment of spatial identifiers (e.g. x, y coordinates, nearest intersection, traffic zone, census block) to activity/trip origins and destinations. Geocoding of survey data is arguably the most challenging aspect of data collection. This is primarily because of inaccuracies and omissions in both the reporting of addresses by survey participants and the address reference databases used in the matching process. In addition, these problems are being exacerbated by the growing demand for survey data with greater spatial precision to support the increasing use of disaggregate approaches in planning studies (Kreitz, 2001).

Geocoding of survey data has traditionally been a tedious, error-prone, manual activity. However, the development of electronic address-matching software and its subsequent incorporation within geographic information system (GIS) packages has led to greater efficiency, accuracy and precision in the geocoding process. While improvements in the quality and comprehensiveness of GIS reference databases have led to further improvements in the geocoding process, the problems of reliance on survey participants for address information remain. In response, a growing number of applications have attempted to use GIS capabilities for prompting/checking location information and automatically assigning geocodes during the data retrieval phase. While the benefits are undoubtedly through integration with automated data collection technologies such as the Global Positioning System (GPS) and sophisticated computer-assisted data recording and retrieval systems that the benefits of GIS technology in travel surveys will be fully realized.

This chapter considers how GIS capabilities have been used to enhance the collection of travel survey data – while it does not focus on the analysis of survey data *per se*, it is clear that effective spatial analysis is only possible if the data are collected and geocoded correctly. The chapter begins by reviewing how GIS have been used in a general sense as part of travel survey methodologies. The next section details how the geocoding process works and how this has been enhanced by the use of GIS capabilities. Following this, the discussion turns to the issues and options associated with developing the two major input requirements for geocoding, the reference and target databases. Finally, some concluding comments are made on where GIS technology might take travel survey methodologies in the future.

## 2. Use of GIS in travel surveys

Figure 1 illustrates the generic travel survey process, which essentially comprises three phases. Following the decision to initiate a survey, the first phase involves planning and designing the survey methodology. This includes the development, testing, and subsequent refinement of the sampling procedures and survey instruments. The second phase covers implementation of the survey, where survey participants are recruited and their travel/activity information is recorded and retrieved. The final phase involves the coding, geocoding, cleaning, and compilation of the data for analysis and presentation.

Through the capability to capture, process, analyze, and display spatially referenced data, GIS offer intuitive appeal to support the process shown in Figure 1. This was recognized in the mid-1990s, when a research synthesis was initiated by the US Department of Transportation to establish how GIS had been used in recent household travel surveys and to provide recommendations on how this technology could be used to enhance future survey methodologies (Greaves, 1997). The synthesis covered 50 household travel surveys conducted in the USA from 1990 to 1997. This period coincided with the rapid growth in GIS capabilities in general, and was reflected in the fact that by the mid-1990s all surveys were making some use of GIS capabilities (compared with only 28% in 1990–1991).

In terms of how GIS technology was being used, this was largely in the third phase of the process shown in Figure 1 for the geocoding and logic checking of data that had already been collected. A few surveys had used GIS capabilities during the data retrieval phase to check/prompt for spatial information and automatically assign geocodes as data were retrieved. The consensus of those surveyed during the synthesis, was that GIS capabilities had significantly reduced the overall processing time and led to improvements in the quality of the spatial data. In addition, the use of spatial overlay tools to analyze and present the data at different levels of aggregation and the ability to integrate data from several

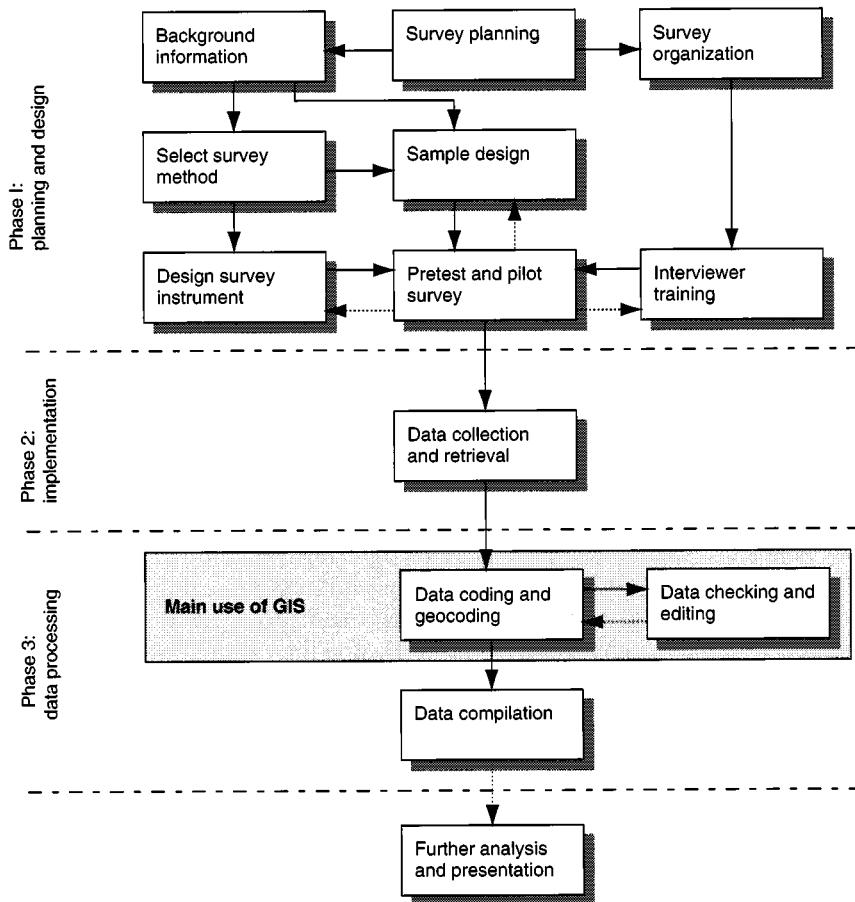


Figure 1. Overview of the travel survey process.

different layers were seen as a major benefit of organizing the survey data within a GIS environment.

### 3. Geocoding of survey data

Geocoding describes the process by which location information is assigned to an address so that it can be spatially referenced. This assignment of geocodes is done through a process known as address matching. Address matching involves the assignment of location information to a target database (i.e. survey trip origins

and destinations) based on a reference database that incorporates address and location information. Until recently, geocoding was largely a manual activity that involved staff locating each address on a map and identifying the appropriate geographic unit into which the address fell. The process was tedious and error-prone, and geographic representation was at best an approximation, with coding usually done to an aggregate area unit such as a census tract or traffic zone.

### *3.1. Automated address matching and GIS*

The development of address-matching software and its subsequent incorporation within GIS software offers many advantages over manual geocoding. These advantages include the ability to geocode to  $x, y$  geographic coordinates, the capability to process entire survey databases without user intervention, the provision of various procedures for dealing with partial matches and rejected records, and the capability to impute the approximate location of missing destinations using visual "heads-up" digitizing tools. While recognizing these potential benefits, it is important to understand how the automated address-matching process works, particularly with regard to how the reference and target database records are structured, how the matching algorithms work, and the various options for dealing with partial matches.

### *3.2. How the automated geocoding process works*

A GIS comprises three data types (features): points, lines and polygons. While survey addresses are considered as point features, they are usually matched to line (e.g. streets) or polygon (e.g. traffic zones, census tracts) features and assigned coordinates through interpolation. The advantage of coding to a coordinate is that this offers flexibility if the geography needs to be redefined such as when zone boundaries are adjusted or data are recoded to different zone systems. Of course, this must be balanced against the fact that coding to a point requires a higher level of accuracy and precision in both the reference and target databases.

The most common form of reference database used for electronic address matching contains street segments, i.e. the section of a street between two intersections. Each segment contains a number of fields, such as street name, street type, directional prefix and suffix, the low and high addresses on both the left and right side of the segment, and other information such as census boundaries and postal (or ZIP) codes. In addition, the latitude and longitude of the segment endpoints may be known, as are optional shape (intermediate) points between the two ends of the segments.

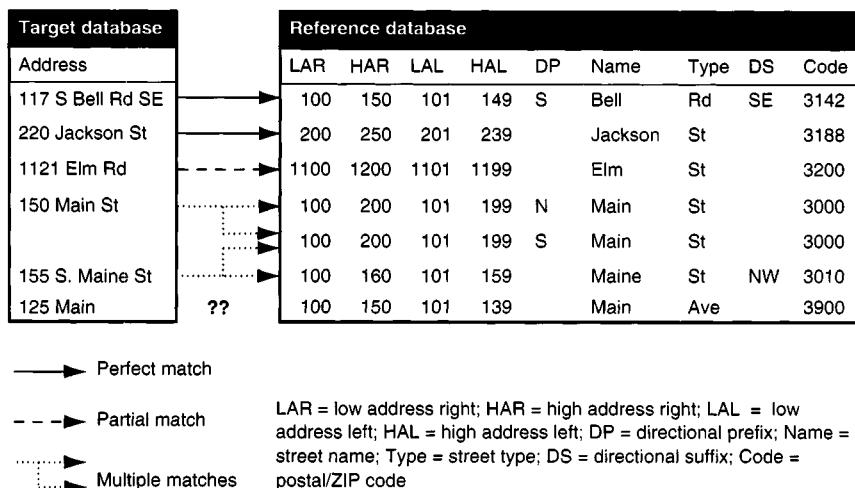


Figure 2. Some typical outcomes of address matching.

Records from the target database are then matched to the appropriate street segment across these identifiers – this is where the importance of probing respondents for comprehensive and accurate address information becomes imperative. In theory, if comprehensive and accurate target and reference databases are prepared, the match rate should be 100%. Unfortunately, this is rarely (if ever) the case. Figure 2 illustrates some common outcomes of the address-matching process, which may result from omissions or ambiguities in the target database. In this case the first two records result in perfect matches since the information is identical in the two databases. The third record results in a partial match, because of the difference in street type. The fourth and fifth records demonstrate cases where partial matches occur with two records, while the final record can only be matched if less stringent matching criteria are employed. It is important to be aware of these issues, particularly when it comes to using automated address-matching procedures.

Once an address has been matched to the correct street segment, the address-matching routine determines which side of the street it should be located on based on the parity (odd/even designation). The next step is to determine the precise location of the address along the segment. Usually, this is done by interpolating the distance along the segment based on the low and high addresses of the segment. Figure 3 shows an example of how this would work for the example of 117 South Bell Street, SE. First, the software would establish that the address was on the left side of the street segment. It would then determine that the address should be located one-third, i.e.  $(117 - 101)/(149 - 101)$ , of the way between

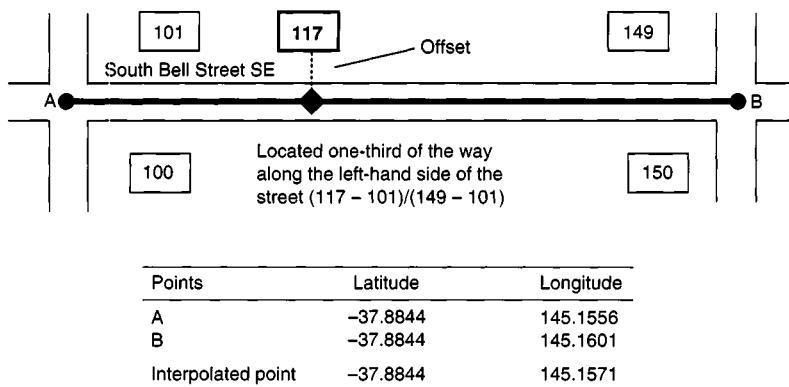


Figure 3. Example of interpolating an address location.

the two end points (A and B) of the segment, and assign coordinates to the interpolated point based on the coordinates of these known points. As a final step, the user will have the option of offsetting the address from the street centerline. This can be done manually or using a predefined offset based on average setbacks of buildings from roadways in the municipality.

Ensuring that the address is on the correct side of the street and offsetting is important not only from the perspective of trying to reflect reality but on how it impacts the use of data at coarser levels of aggregation. If the street segment forms the boundary between two areal units of aggregation such as traffic zones, it is critical the data point (1) does not sit on the boundary, and (2) is in the correct zone.

### 3.3. Partial matches

According to one source, automated address matching typically results in anywhere from 25 to 75% of addresses being successfully matched (Drummond, 1995). While these percentages can be improved with diligence in the preparation of the reference and target databases, clearly procedures must be implemented to deal with rejected records. Several techniques have been developed by GIS vendors to improve match rates, although it must be realized these essentially trade off the achievement of a higher proportion of successful matches with the potential for error. Some of the most common techniques are described here.

Many GIS have the capability to parse addresses; that is, to interpret the survey record in such a way that it can be matched to the database. Typical address-parsing features include standardization of common elements (e.g. "Avenue,"

“Ave.” and “Av” as “Ave”), automatic creation and identification of address fields, methods for dealing with vague or ambiguous addresses, and surrogates for common street names, allowing them to be referred to in different ways (e.g. Martin Luther King Jr Blvd, M.L. King Blvd, MLK Blvd, etc.). An additional parsing feature of significance in the context of survey data is the capability to parse intersection addresses because many participants are able to relate a location to the nearest intersection rather than an exact address.

In addition to parsing capabilities, many GIS packages include a variety of transformation options and procedural strategies designed to improve the match rate while keeping the user informed about the impacts on the match rate. Criteria relaxation (e.g. the MapInfo software package) involves multiple passes through the database, with a successive loosening of the requirements for a match. Rating procedures (e.g. the TransCAD software package) enable the user to define the strictness of a match and review potential matches. The scoring table approach (e.g. the ArcView software package) involves the application of a soundex function to the street name followed by the generation of a list of candidate street segments. A soundex function dramatically increases the potential match rate by assigning a phonetic equivalent for the written spelling of a street name (e.g. “Maine,” “Main” and “Mane” might all have the same soundex, “MN”). Each candidate segment begins with a score of 100, with points deducted for each non-matching element. The one with the highest score is then assigned the match.

### *3.4. Checking of geocodes*

Whatever procedure is used, it is critical to realize the potential for error increases as the matching criterion are loosened. The implications are that while the geocoding itself may progress much quicker than previously, the checking of geocodes can still be a time-consuming and tedious task. Among the most usual checks of geocodes are:

- aggregation checks on the location of geocodes – very often when using automated procedures, systematic errors only become apparent when aggregating the data to a zone level;
- checking against other information such as telephone exchanges – this is particularly critical in large metropolitan areas where several instances of the same street name occur;
- verification that one trip began where the previous trip ended – i.e. ensuring the correct linkage between reported trips;
- cross-checking distances and times calculated from geocoded points with participant-provided information – this is another method to pick up very apparent errors based on large discrepancies.

#### 4. Developing the databases

The most critical and problematic aspect of geocoding is the development of accurate and complete target and reference databases. This section considers some of the major issues in more depth.

##### 4.1. *The reference databases*

The options for and the quality of the automated geocoding process are dictated to a large extent by the availability of comprehensive, accurate and up-to-date electronic reference databases. Reference databases range from simple digitized boundaries of area units of census geography to digital street centerline files that support coordinate geocoding of addresses using procedures similar to those described previously. These databases may be developed locally or purchased from one of a number of commercial vendors.

In the USA, the most widely used source for developing electronic address-matching databases is the street centerline file from the Topologically Integrated Geographic Encoding and Referencing (TIGER/Line) system. These files are part of a spatial support system developed by the US Bureau of the Census for its census and survey programs. They include political and statistical geographic areas, address ranges, and ZIP codes, and are freely available for download in a GIS-friendly format from [www.esri.com/data/download/census2000\\_tigerline/index.html](http://www.esri.com/data/download/census2000_tigerline/index.html). Problems and issues with using TIGER/Line files are well known and documented, particularly with respect to the poor coverage of non-residential and non-urban locations, missing/erroneous records, their currentness, and spatial inaccuracy stemming from how they were originally digitized (Greaves, 1997).

Many agencies have taken the TIGER/Line files and progressively improved accuracy and coverage for their region. More problematic has been how to develop and maintain employment and other non-residential databases. Invariably, this involves either a substantial in-house effort or the purchase of such a database from a commercial vendor. A promising option for major enhancements to electronic reference databases is the digitization of property boundaries (often referred to as parcels) in response to the needs of agencies that require precise address details (e.g. emergency services and property assessors). A parcel database typically contains property limits, address information, and the establishment name, and offers greater accuracy than a street segment file because the *x*, *y* coordinates can be interpolated from the parcel boundaries – the simplest approach is to take the centroid of the parcel.

Whatever option is selected, it is rare that the reference database has been developed specifically for the geocoding of travel survey data. This implies that some pre-processing of data is invariably required to try to “match” the ways in

which participants report address information. Typical measures include the standardization of address elements, the incorporation of colloquialisms, the inclusion of alternative street names, and (as apparent as it may seem) the inclusion of the establishment name.

#### *4.2. Developing the target database*

Ideally, one would like to collect the street number, street direction prefix, street name, street suffix, city, and zip code plus the place of business establishment name for each location.

(Stecher et al., 1995)

The difficulties with developing spatially accurate and precise travel survey databases emanate from two primary causes. The first, and most significant, is that other than frequently visited locations (e.g. the workplace and school), participants either forget or cannot provide sufficient detail for geocoding on the places they visit during the survey period. This problem has generally been exacerbated by the use of GIS, which while undoubtedly providing the potential to improve the quality of spatial travel data require more precise address information to take advantage of the enhanced geocoding capabilities. The second cause stems from the timeliness of the geocoding and the fact it often occurs months after the data are collected, so ambiguous information cannot be clarified by the participants. The apparent resolution to these problems is to probe for, check, and even code the required address information during the data retrieval (and potentially data collection) phase to cut down on the problems with working in a post-processing mode. GIS technology has the potential to markedly assist with this process, particularly through integration with increasingly sophisticated electronic data capture and retrieval systems. Some of the more recent and innovative developments are discussed here.

#### *Computer-assisted telephone interview*

Many travel surveys now retrieve information using a centralized telephone system. This has been driven by (1) the higher unit costs and concerns over personal safety associated with face-to-face interviews, and (2) the development of computer-assisted telephone interview (CATI) systems to automate many of the data retrieval activities that were traditionally done manually (Greaves, 1997). CATI systems are typically customized applications that include features such as menus and built-in logic checks to only permit entry of legitimate codes, cross-checks of the consistency of data with previously entered data, elimination of redundant questions, customization of question sequencing based on responses to previous questions, prompts for information, and coding of survey data.

Intuitively, checking and geocoding of spatial data could also be incorporated within the CATI process. However, this has proven particularly complex because it relies on highly accurate geographic files that can be referenced based on participant responses and skilled interviewers able to probe for information until an address is found without extending the interview time to unacceptable levels.

One of the first efforts to incorporate spatial checking/prompting capabilities as part of the CATI retrieval system was in a household travel survey of the Greater Toronto Area in 1991 (Ng and Sarjeant, 1993). Here, in addition to many of the features described previously, the customized CATI system incorporated (1) look-up tables of public transportation routes that enabled trips to be verified and coded as the interview continued, (2) look-up tables of school names, municipalities, and street names to assist the interviewer in prompting for correct information, and (3) a spelling tool that enabled the interviewer to type the first few letters of an establishment name, which would then bring up all the possible matches and could be used to prompt the participant for the correct location. The authors report that although the costs of software development were high, the savings in post-processing time more than compensated for the initial expense.

A more recent effort to integrate GIS capabilities within a CATI retrieval system was used in the Hawaiian household travel survey of 1996. The Hawaiian language incorporates 13 vowels, which caused particular confusion when describing locations. The solution was to code the street file and major landmarks into the CATI system, to assist the interviewers in prompting participants for the correct spelling of addresses during recall. Another approach was employed in the recently completed New York metropolitan survey, which presented particular challenges due to the sheer size and complexity of the region. Given that it was impractical to try to code a street file within the CATI system, the solution was for the interviewer to have two screens: one with a CATI-style interface and the other with a graphical display of the region that was used to visually verify locations. While no reports of the merit of this approach have been made available at the time of writing, this approach had major benefits during the pilot study of the late 1990s.

Few attempts have been instigated to actually assign geocodes "on-line" during the CATI data retrieval. Baber and Bandy (1995) describe one attempt to incorporate this capability during the Baltimore household travel survey of 1993, using a customized CATI system that incorporated an enhanced emergency service database and address-matching software. The "on-line" geocoding was actually done to the 1990 traffic zone system and was recoded to longitude-latitude at a later date. This proved to be a time-consuming and problematic task, with approximately 80% of addresses successfully geocoded in this manner.

### *Computer-assisted personal interview*

While computer-assisted techniques are seemingly necessary to capture the increasingly complex behavioral and spatial activity data required from travel surveys, opinion is divided over the most appropriate way to record and retrieve these data. While CATI systems predominate in large surveys in the USA for the reasons cited previously, the preclusion of face-to-face contact has led to both a decline in the response rates and more difficulty in retrieving the data. In addition, CATI retrieval continues to suffer from the fact that certain segments of the population are increasingly difficult to reach by telephone and increasingly resistant to telephone surveys in general. Face-to-face interviews have generally been more prevalent for targeted surveys of “rarer” populations such as transit and bicycle users, where it is more efficient to sample people in a context (e.g. on board) rather than rely on a household survey.

From the perspective of spatial data, personal interviews offer unquestionable advantages because the interviewer can prompt the respondent for information with maps and other supplementary material. The use of graphical means to assist with recalling location information is not a new concept. For example, in the 1980s, as part of the Household Activity Travel Simulator (HATS) project in the UK, respondents were interviewed at home, and maps were used to try to ascertain how participants would respond to certain policy changes (Jones, 1980). As technology has progressed, the development of computer-assisted personal interview (CAPI) systems that integrate GIS capabilities has become increasingly popular. For instance, Flood and Siaurusaitis (1997) describe one such application where respondents used on-site computerized touch-screen maps to pinpoint their trip origins and destinations. A similar approach is described in Wermuth et al. (2001), although they note this method of interview is relatively expensive when compared with other computer-assisted options.

### *Computer-assisted self interview*

Another approach being actively pursued involves the development of computer-assisted self-interview (CASI) techniques. These techniques have evolved from traditional paper-and-pencil techniques and involve direct interaction between the participant and a computer, palmtop, or mobile phone (Wermuth et al., 2001). Customized applications are integrated within these devices to lead the participant through the survey and to record the relevant data directly.

CASI applications are becoming increasingly sophisticated, and many now include graphical and GIS capabilities to prompt for and record location information. For example, Kreitz (2001) describes the addition of GIS functionality to the Computerized Household Activity Scheduling Elicitor (CHASE), originally developed by Doherty et al. (1999). The CHASE system works like an electronic

day planner, with participant-provided information on planned activities fed into the system at the beginning of the week. Activities can be added or modified to the system as the week progresses. CHASE-GIS includes landmarks, road characteristics, and public transport maps, and serves as both a display tool and a spatial data input tool, with geocoding taking place automatically. Activity locations can be chosen or added by clicking on the map. The system then automatically selects proposed routes, which the user can modify if they do not match reality.

Another development in CASI systems is being facilitated by the rapid expansion in the use and capabilities of the Internet. While issues of selection bias are clearly a concern, the appeal of the Internet is that it essentially combines the functionality of CATI systems with the graphic capabilities of CASI systems on the backbone of powerful Internet servers (Adler et al., 2002). Computer-assisted Web interviews (CAWIs) are commonplace, and with their low deployment costs and the increasing problems of using home telephones for surveys are seemingly an inevitable component of the future of travel surveys. Adler et al. detail one application in which the participants are able to specify locations based on (1) a street address, (2) the nearest intersection, (3) a business name, (4) map point and click, (5) an external location, or (6) a previously identified location. The system incorporates all the consistency and logic checks of CATI, and then assigns geocodes to the location information.

### *Integration of GIS with automated data collection techniques*

While these techniques all increase the likelihood of improving the quality of spatial travel data, the continued reliance on self-reporting is troublesome. Arguably, what is required is a method for automatically recording where participants engage in activities to either verify the information being provided or to derive it directly. Since the concept was first proven in Lexington, Kentucky (Murakami and Wagner, 1996), the notion and advantages of using GPS technology to record spatial elements of travel have expanded worldwide. While the use of GPS technology for data collection is dealt with in detail in Chapter 19, its inextricable link with GIS deserves some attention here.

GPS provides time-stamped data on latitude and longitude, meaning it can be directly referenced into a GIS environment (with the appropriate translation dependent on the spatial datum being used). Through map matching to digital street centerline files and the employment of rules about what constitutes a stop, it is possible to determine trip end location, the routes used, travel times, and travel distance (Stopher and Bullock, 2001). One example of the types of plot possible with GPS travel data using a GIS is shown in Figure 4. In addition, it is now also possible to derive previously unmeasurable vehicle-operating characteristics (e.g. speeds, accelerations, decelerations, and idle times) as part of the collection of

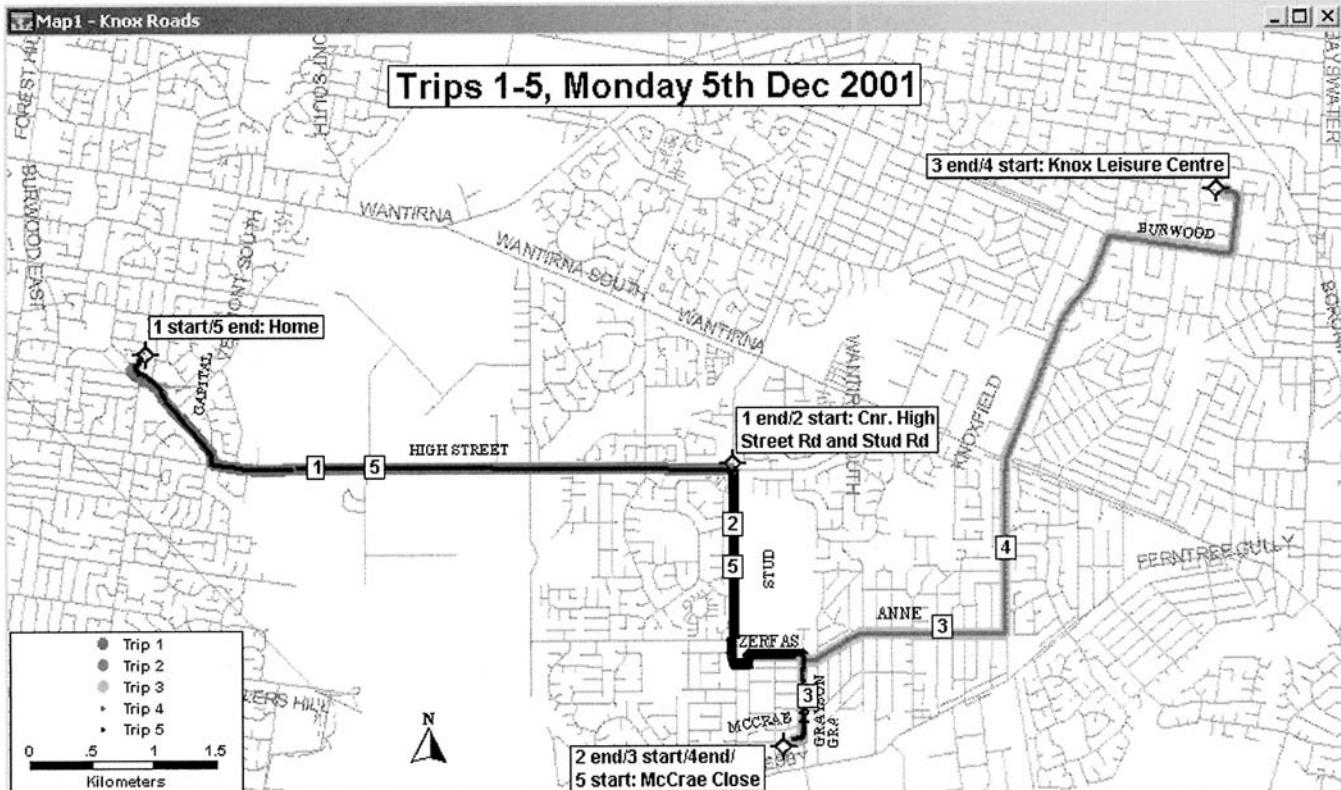


Figure 4. Displaying GPS-recorded routes in a GIS environment.

survey data and to plot these on a map to show the spatial distribution of such episodes (Greaves and De Gruyter, 2002).

While GPS provides location information, the question of how to derive other critical information, such as the reasons for travel and number of vehicle occupants, remain. Two schools of thought have emerged on this issue. The first involves enhancements to CASI systems to include automatic GPS data logging of position. This was the procedure followed in the original Lexington study, and is currently being pursued as an enhancement to the CHASE system described previously (Doherty et al., 1999).

The second school of thought takes the view that the primary objective should be to reduce (or even eliminate) the need for participant intervention in the recording of the actual field data (Stopher and Bullock, 2001). Under this approach, GPS data are recorded and then processed within a GIS, to prompt respondents for other critical information on their trips (e.g. purpose and vehicle occupants). In a proof-of-concept test of this approach, Bachu et al. (2001) found that the time taken for prompted recall utilizing GPS was 15–20% of that required for reporting in a telephone interview. Wolf et al. (2001) demonstrate that it could be possible to reduce participant input still further by imputing trip purpose based on GPS data processed within a highly accurate and comprehensive GIS electronic database.

#### *4.3. Spatial bias and spatial stratification*

GIS technology also has two potentially important roles to play in relation to the spatial representation of the survey participants. The first is in relation to the stratum used to recruit participants. Currently, other than some coarse spatial delineation based on political units (e.g. counties or districts), travel survey participants are rarely recruited according to land use type. This makes a reliable assessment of the impact of land use policies on travel patterns somewhat problematic. GIS technology (potentially) provides the capability to delineate specific land use stratum for recruiting participants. For example, the 1994–1995 Oregon–SW Washington travel survey wanted to ensure adequate representation from different modes of travel, particularly public transport, bicycling, and walking. “Market areas” were developed for sampling from the downtown area of Portland, and delineated through a GIS, based on different urban design and public transport criteria. In addition to being popular for mode-specific studies, this type of approach is also appealing for corridor-planning studies, where some “proximity to the corridor” measure could be useful.

The second role of GIS in this context is to identify any problems of spatial bias with the sample. Spatial bias involves the under- or over-representation of participants from particular geographic areas, which may result from the limited

geographic context of sampling frames (Greaves, 1997). For instance, the 1993 Albuquerque household travel survey used a published telephone list to draw the sample. In plotting out the locations of participants, serious spatial biases were uncovered, because it transpired that the listed numbers were predominantly in one region of the city while unlisted numbers were in another; call patterns were adjusted accordingly. This type of checking is even more useful if done during the conduct of the survey so that call adjustments can be made accordingly.

## 5. Summary and future directions

Advanced technologies are increasingly shaping the way that travel data are collected, processed, and analyzed. This chapter has focused on one such technology, GIS, and demonstrated how it has enhanced the travel survey process from automating traditionally tedious tasks such as geocoding to potentially changing the way surveys are designed, organized and implemented.

It is important to stress that while the sophisticated technology may appeal, the benefits of GIS do not come without considerable time and effort. This includes the development of comprehensive and spatially accurate reference databases, which are troublesome and expensive to create. Invariably, these databases will be developed – many agencies now have GIS departments with staff working full time on their creation. Even then, the databases are rarely contiguous with how survey respondents provide address information, which implies careful database parsing is required. This is even more imperative when these databases are integrated with the types of applications described in Section 4.

Referring back to Figure 1, it was noted that in the synthesis conducted in the mid-1990s, the predominant use of GIS was to code and check data that had already been collected. While this is still largely the case, it appears the use of GIS will continue to grow in other elements of the travel survey process. In terms of sample design, it is seemingly important to consider spatial location when sampling because of the difficulty of answering complex land use/travel implication questions using current data sources. GIS technology is intuitively suited for this task through the possibility to define “market areas” based on spatial accessibility measures, and land use descriptors.

In the area of data processing, clearly, many subsequent problems can be mitigated or removed by incorporating these processing tasks during data collection itself (i.e. in phase 2). The development of computer-assisted data collection devices with GIS capabilities is facilitating this, and rapid developments can be expected in the future, particularly in the area of CASI and CAWI techniques. Further enhancements are probable, with the continued development of automated data collection technology such as GPS and GSM (Global System of Mobile Communication). Importantly, these techniques are seemingly able

to provide the required travel information without overburdening survey participants. In addition, and arguably most important of all, they may provide a way to re-ignite peoples' willingness to participate in travel surveys by simply making them more enjoyable!

## References

- Adler, T., L. Rimmer and D. Carpenter (2002) "Use of an Internet-based household travel diary survey instrument," in: *Proceedings of the 81st Annual Meeting of the Transportation Research Board*. Washington, DC.
- Baber, C.M. and G. Bandy (1995) "Baltimore region household survey: methods and procedures," in: *Proceedings of the 5th National Conference on Transportation Planning Applications*, Vol. 1. Seattle.
- Bachu, P.K., T. Dudala and S. Kothuri (2001) "Prompted recall in global positioning system survey: proof of concept study," *Transportation Research Record*, 1768:106–113.
- Doherty, S.T., N. Noel, M. Gosselin, C. Sirois and M. Ueno (1999) "Moving beyond observed outcomes: integrating global positioning systems and interactive computer-based travel behaviour surveys," in: *Personal travel: the long and short of it*. *Transportation Research*, E-Circular 026:449–466.
- Drummond, W.J. (1995) "Address matching: GIS Technology for mapping human activity patterns," *Journal of the American Planning Association*, 61/62:240–251.
- Flood, M. and V. Siaurusaitis (1997) "GIS spatial analysis tools for traffic model inputs," in: *Proceedings of the 76th Annual Meeting of the Transportation Research Board*. Washington, DC.
- Greaves, S.P. (1997) "Applications of GIS technology in recent household travel survey methodologies," in: *Proceedings of the GIS-T Symposium*. Greensboro.
- Greaves, S.P. and C. De Gruyter (2002) "Profiling driving behaviour using passive Global Positioning System (GPS) technology," in: *Proceedings of the Institute of Transportation Engineers International Conference*. Melbourne.
- Jones, P. (1980) "Experience with household activity-travel simulator (HATS)," *Transportation Research Record*, 765:6–12.
- Kreitz, M. (2001) "Methods for collecting spatial data in household travel surveys," in: *International Conference on Transport Survey Quality and Innovation*, Plenary Paper. Kruger National Park.
- Murakami, E. and D.P. Wagner (1996) "Global Positioning Systems for personal travel surveys," in: *NATDAC '96*, Paper. Albuquerque.
- Ng, J.C. and P.M. Sarjeant (1993) "The use of direct data entry for travel surveys," in: *Proceedings of the 72nd Annual Meeting of the Transportation Research Board*. Washington, DC.
- Stecher, C., S. Bricka and L. Goldenberg (1995) "Travel behaviour survey data collection instruments," in: *Conference on Household Travel Surveys: New Concepts and Research Needs*, Resource Paper. Irvine.
- Stopher, P.R. (2000) "Survey and sampling strategies," in: D.A. Hensher and K.J. Button, eds, *Handbook of transport modelling*. Oxford: Pergamon.
- Stopher, P.R. and P. Bullock (2001) "Using passive GPS as a means to improve spatial travel data: further findings," in: *Proceedings of the 23rd Conference of the Australian Institutes of Transport Research*. Melbourne.
- Wermuth, M., C. Sommer and M. Kreitz (2001) "Impact of new technologies in travel surveys," in: *International Conference on Transport Survey Quality and Innovation*, Plenary Paper. Kruger National Park.
- Wolf, J., R. Guensler and W. Bachman (2001) "Elimination of the travel diary: experiment to derive trip purpose from Global Positioning System travel data," *Transportation Research Record*, 1768:125–134.

*Chapter 22*

## GIS AND NETWORK ANALYSIS

MANFRED M. FISCHER

*Vienna University of Economics and Business Administration*

### 1. Introduction

Both geographic information systems (GIS) and network analysis are burgeoning fields, characterized by rapid methodological and scientific advances in recent years. A GIS is a digital computer application designed for the capture, storage, manipulation, analysis and display of geographic information. Geographic location is the element that distinguishes geographic information from all other types of information. Without location, data are termed to be non-spatial and would have little value within a GIS. Location is, thus, the basis for many benefits of GIS: the ability to map, the ability to measure distances, and the ability to tie different kinds of information together because they refer to the same place (Longley et al., 2001).

GIS-T, the application of geographic information science and systems to transportation problems, represents one of the most important application areas of GIS technology today. While the strengths of standard GIS technology are in mapping display and geodata processing, GIS-T requires new data structures to represent the complexities of transportation networks and to perform different network algorithms in order to fulfil its potential in the field of logistics and distribution logistics.

This chapter addresses these issues as follows. The section that follows discusses data models and design issues that are specifically oriented to GIS-T, and identifies several improvements of the traditional network data model that are needed to support advanced network analysis in a ground transportation context. These improvements include turn-tables, dynamic segmentation, linear referencing, traffic lines and non-planar networks. Most commercial GIS software vendors have extended their basic GIS data model during the past two decades to incorporate these innovations (Goodchild, 1998).

The third section shifts attention to network routing problems that have become prominent in GIS-T: the traveling-salesman problem, the vehicle-routing problem and the shortest-path problem with time windows, a problem that occurs as a

subproblem in many time-constrained routing and scheduling issues of practical importance. Such problems are conceptually simple but mathematically complex and challenging. The focus is on theory and algorithms for solving these problems. The chapter concludes with some final remarks.

## 2. Network representation and GIS-T network data models

### 2.1. Terminology

A network is referred to as a pure network if only its topology and connectivity are considered. If a network is characterized by its topology and flow characteristics (such as capacity constraints, path choice, and link cost functions) it is referred to as a flow network. A transportation network is a flow network representing the movement of people, vehicles, or goods (Bell and Iida, 1997).

The approach adopted almost universally is to represent a transportation network by a set of nodes and a set of links. The nodes represent points in space and possibly also in time, and the links tend to correspond to identifiable pieces of transport infrastructure (e.g. a section of road or railway). Links may be either directed, in which case they specify the direction of movement, or undirected.

In graph theoretical terminology, a transportation network can be referred to as a valued graph, or alternatively as a network. Directed links are referred to as arcs, while undirected links are edges. Other useful terms with some intuitive interpretations are a path, which is a sequence of distinct nodes connected in one direction by links; a cycle, which is a path connected to itself at the ends; and a tree, which is a network where every node is visited once and only once. The relationship between the nodes and the arcs, referred to as the network topology, can be specified by a node–arc incidence matrix: a table of binary or ternary variables stating the presence or absence of a relationship between network elements. The node–arc incidence matrix specifies the network topology, and is useful for network processing.

### 2.2. The network data model

The heart of any GIS is its data model. A data model is an abstract representation of some real-world situation used to organize data in a database. Data models typically consist of three major components. The first is a set of data objects or entity types that form the basic building blocks for the database. The second component is a set of general integrity rules that constrain the occurrences of entities to those that can legally appear in the database. The final component

includes operators that can be applied to entities in the database (Miller and Shaw, 2001).

Data modeling involves three different levels of abstraction: conceptual, logical, and physical. Conceptual data models describe the organization of data at a high level of abstraction, without taking implementation aspects into account. The entity–relationship and the extended entity–relationship models are the most widely used conceptual data models. They provide a series of concepts such as entities, relationships, and attributes, capable of describing the data requirements of an application in a manner that is easy to understand and independent of the criteria for managing and organizing data on the system. A logical data model translates the conceptual model into a system-specific data scheme, while low-level physical data models provide the details of physical implementation (file organization and indexes) on a given logical data model (Atzeni et al., 1999).

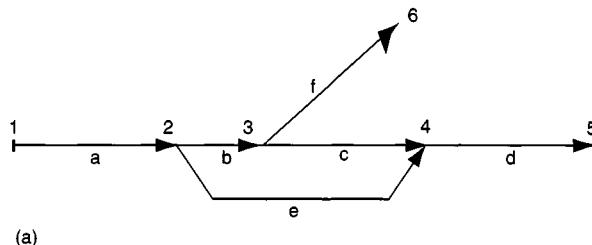
The network data model is the most popular conceptual model to represent a network within a GIS environment. The model – a special type of the node–arc–area data model that underlies many basic vector GIS databases – is built around two core entities: the node (a zero-dimensional entity) and the arc (a one-dimensional entity). Current GIS data models typically represent a network as a collection of arcs with nodes created at the arc intersections. The planar embedding of the node–arc data model guarantees topological consistency of the network.

The most widely used logical data model that supports the node–arc representation of networks is the georelational model. This model separates spatial and attribute data into different data models. A logical spatial data model (the vector data model) that encodes nodes and arcs maintains the geometry and associated topological information, while the associated attribute information is held in relational database management system (RDBMS) tables. Unique identifiers associated with each spatial entity (node or arc) provide links to records in the relational model and its data on the attributes of the entity. This hybrid data management strategy was developed to take advantage of an RDBMS to store and manipulate attribute information (Longley et al., 2001). But this solution does not allow the relationships between a spatial object and its attributes to have their own attributes (Goodchild, 1998). Though the solution is neither elegant nor robust, it is effective, and the georelational model is widely present in GIS software (Miller and Shaw, 2001).

The relational structure to support the planar network model typically consists of an arc relation and a node relation. The structure may be illustrated as a representation of the simple network shown graphically in Figure 1a. The model implemented in GIS represents each arc of the network as a polyline entity. Associated with each entity will be a set of attributes, conceived as the entries in one row of a rectangular table (Figure 1b). Properties may include information about the transverse structure such as the number of lanes or information on

address locations within the network. Commonly included attributes are arc length, free flow travel time, base flow, and estimated flow. The base and estimated flows usually refer to the observed flow and the flow estimated from some modeling exercise (Miller and Shaw, 2001).

The node relation typically contains a node identification field and relevant attributes of the node, such as the presence of a traffic light. Figure 1d shows a scheme that includes storage of pointers from arcs to nodes, and from nodes to arcs, to store the topology of the network (connectivity). Each arc has an inherent direction defined by the order of points in its polyline representation, as illustrated by arrows in Figure 1a (Goodchild, 1998). It is noteworthy that the



(a)

Arc ID	Street name	Lanes	Other attributes
a	High Street	2	
b	High Street	4	
c	High Street	4	
d	High Street	2	
e	River Way	2	
f	Hill Street	2	

(b)

Node ID	Stop light	Other attributes
1	n	
2	y	
3	n	
4	y	
5	n	
6	n	

(c)

Arc ID	Street name	Lanes	From node	To node
a	High Street	2	1	2
b	High Street	4	2	3
c	High Street	4	3	4
d	High Street	2	4	5
e	River Way	2	2	4
f	Hill Street	2	3	6

(d)

Node ID	Stop light	Arc links
1	n	a
2	y	a, b
3	n	b, c, f
4	y	c, d, e
5	n	d
6	n	f

Figure 1. Relational data model representations of the arcs and nodes of a network. (a) Example network for the relational model example. (b) A simple arc table. (c) A simple node table. (d) Pointers added to the arc and node tables to represent connectivity. (Adapted from Goodchild, 1998.)

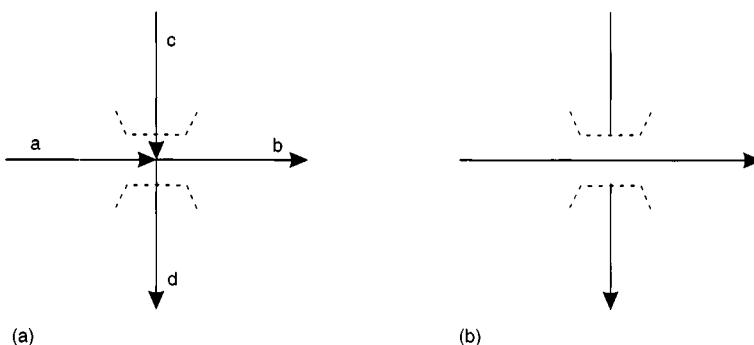


Figure 2. Illustration of (a) planar and (b) non-planar network approaches to represent an overpass.

node–arc representation disaggregates a transportation system into separate subnetworks for each mode within a single base network. Transfer arcs link the subnetworks. The pseudo-arcs represent modal transfers.

### 2.3. Non-planar networks and the turn-table

The planar network data model has received widespread acceptance and use. Despite its popularity, the model has limitations for some areas of transportation analysis, especially where complex network structures are involved. One major problem is caused by the planar embedding requirement. This requirement forces nodes to exist at all arc intersections and, thus, ensures topological consistency of the model. But intersections at grade cannot be distinguished from intersections with an overpass or underpass that do not cross at grade. This difficulty in representing underpasses or overpasses may lead to problems when running routing algorithms (Kwan et al., 1996). The drawbacks in planar topology for network representations have motivated interest in non-planar network models. Planar models force nodes at all intersections (Figure 2a), while non-planar network models (Figure 2b) do not. Non-planar networks are broadly defined as those networks that permit arcs of the network to cross without a network node being located at the intersection. There is no implicit or explicit contact between the line segments at the point of intersections (Fohl et al., 1996).

But non-planar data models provide only a partial solution to the problem of connectivity. In transportation network analysis it may be necessary to include extensive information on the ability to connect from one arc to another. Drivers, for example, may force turn restrictions, or trucks may be limited by turning radius. Such situations require more than the simple ability to represent the existence of a crossing at grade or an underpass (Goodchild, 1998).

To resolve this problem, the standard fully intersected planar network data model has been extended by adding a new structure, called the turn-table. Table 1 shows a turn-table for the layout used in Figure 2a. For each ordered pair of arcs incident at a node, a row of attributes in the table gives appropriate characteristics of the turn (yes/no), together with links to the tables that contain the attributes of the arcs. In this way, a data model with a planar embedding requirement can represent overpasses and underpasses by preventing turns (Goodchild, 1998).

#### *2.4. Linear referencing systems and dynamic segmentation*

While geographic features are typically located using planar referencing systems, many characteristics associated with a transportation network are located by means of a linear rather than coordinate-based system. These characteristics include data on transportation-related events and facilities (often termed feature data). In order to use linear-referenced attributes in conjunction with a spatially referenced transportation network, there must be some means of linking the two referencing systems together (Spear and Lakshmanan, 1998).

Linear referencing systems typically consist of three components (Vonderohe and Hepworth, 1996; Sutton, 1997): a transportation network, a linear referencing method, and a datum. The transportation network is represented by the conventional node–arc network. The linear referencing method determines an unknown location within the network using a defined path and an offset distance along that path from some known location (Miller and Shaw, 2001). The datum is the set of objects (so-called reference or anchor points) with known georeferenced

Table 1  
Layout of a turn-table for the layout used in Figure 2a

From arc	To arc	Turn?
a	c	n
a	b	y
a	d	n
b	a	y
b	c	n
b	d	n
c	a	n
c	b	n
c	d	y
d	a	n
d	b	n
d	c	y

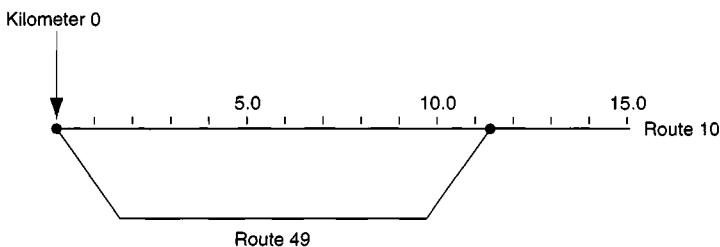


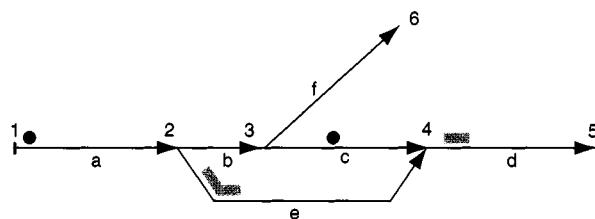
Figure 3. Kilometerpoint referencing.

locations that can be used to anchor the distance calculations for the linear referenced objects.

There are different linear referencing methods. Nyerges (1990) identifies three major strategies, namely road name and kilometerpoint (milepoint) referencing, control section locational referencing, and link and node locational referencing. Road name and kilometerpoint is a system familiar to anyone who has driven on highways in Europe or the USA. This system consists of a road-naming convention (i.e. a standard procedure for assigning names to highways and streets) and a series of kilometerpoint references (i.e. distance calculations along the network, typically measured in fractions of a kilometer or mile). Kilometerpoint referencing requires a designated point of reference (e.g. a kilometer 0) as a datum (Figure 3). This is often an end point of the route or where the route crosses a provincial or a national boundary (Miller and Shaw, 2001).

Owing to road modifications and other changes in road geometry, kilometer-point referencing can become increasingly inaccurate over time. In other words, the reference kilometerpoint may not reflect the actual distance from the point of origin. This may cause problems when maintaining historical records of transportation events, and requires some type of translation factor to adjust distances (Nyerges, 1990; Miller and Shaw, 2001).

The key to tie (zero-dimensional and one-dimensional) objects located at arbitrary locations on the network to the node–arc structure of the network data model is dynamic segmentation. The term derives from the fact that feature data values are held separately from the actual network route in database tables and then dynamically added to segments of the route each time the user queries the database (Longley et al., 2001). Several commercial GIS software packages provide dynamic segmentation capabilities, typically maintained at the logical level using the relational data model (Miller and Shaw, 2001). Figure 4 provides a simple illustration of the concept, for two types of objects located at arbitrary locations on the network. These entities – termed network points (point events) and network segments (line events) – are given their own attribute tables. Dynamic segmentation reduces the number of transportation features or network



- Bus depot (network point)
- Congested route (network segment)

Arc	Distance from arc start	Feature
a	0.7	Bus depot
c	2.2	Bus depot

(a)

Arc	Distance from arc start to start of feature	Distance from arc start to end of feature	Level of congestion
d	0.5	1.2	High
e	1.0	2.5	High

(b)

Figure 4. The concept of dynamic segmentation: a simple example for (a) network points and (b) network segments. (Adapted from Goodchild, 1998.)

links that have to be maintained to represent the system, and is particularly useful in situations where the event data change frequently and need to be stored in a database for access by other applications (Longley et al., 2001).

## 2.5. Lanes and navigable data models

A straightforward way to enhance the basic node–arc model for intelligent transportation systems (ITS) applications is to add information on the transverse structure of the network (Miller and Shaw, 2001). Even though certain information about the transverse structure (such as the existence of a median or the number of lanes) might be stored as attributes of arcs or network lines, it is not possible to store detailed information about individual lanes or connectivity at the lane level. There is, for example, no way to disaggregate a turn-table to store turn restrictions that are specific to lanes (Goodchild, 1998).

ITS database requirements go well beyond the traditional requirements of maintaining arc–node topology, two-dimensional georeferencing, and linear referencing of events within transportation networks. A fully fledged ITS requires a high-integrity, real-time information system that can receive inputs from sensors embedded within transportation facilities and from vehicles equipped with Global

Positioning System (GPS) devices and navigable data models. Navigable data models are digital geographic databases of a transportation system that can support vehicle guidance operations of different kinds. For intelligent transportation systems, this includes four functions (Dane and Rizos, 1998; Miller and Shaw, 2001)

- The data model has to unambiguously translate coordinate-based locations into street addresses, and vice versa. Travelers utilize address systems for location referencing, while an ITS tracks a vehicle utilizing a GPS receiver that can provide locations at accuracies of 5–10 m.
- The data model has to support map matching. This refers to the ability to snap the position of vehicle to the nearest location on a network segment when its estimated or measured location is outside the network. This may occur due to differences in accuracy between the digital network database and the GPS system.
- The data model has to have the capability to represent the transportation network in detail sufficient to perform different network algorithms, modeling, and simulations. In the real world, a transportation network has different types of intersections that are of interest to ITS builders. For some applications, information on intersections, lanes and lane changes, highway entrances and exits, etc., is important. Other applications may require geometric representation of road curvature and incline.
- The data model must not only assist the traveler in selecting an optimal route based on stated criteria such as travel time, cost and navigational simplicity but also support route guidance. This refers to navigational instructions, and is a challenging task in real time.

Although dynamic segmentation can be used to enhance the traditional node–arc structure for ITS applications, much of the high-resolution positional information provided by in-vehicle GPS receivers is lost when referenced within the traditional network structure. While 50 m accuracy may be sufficient to locate a vehicle on a road, better than 5 m will be required to locate to the lane level. Such accuracies are well beyond the capability of many of the currently available network databases. Achievement of better than 5 m accuracy using GPS requires the use of differential techniques and a high level of geodetic control (Goodchild, 1998).

Fohl et al. (1996) describe a prototype lane-based navigable data model where each lane is represented as a distinct entity, with its own connectivity with other lanes, but its geometry is obtained from the standard linear geometry of the road. No attempt is made to store the relative positions of lanes, but the structure does identify such topological properties as adjacency, and the order of lanes across the road (Goodchild, 1998). A more radical approach to navigable data models for ITS is to abandon the node–arc model entirely. Bespalko et al. (1998) suggest

a three-dimensional object-oriented GIS-T data model that can distinguish between overpasses, underpasses, and intersections, thereby providing guidance through complex intersections.

### 3. Vehicle routing within a network: problems and algorithms

At the core of many procedures in GIS-T software are algorithms for solving network-routing problems. The problems are conceptually simple but mathematically complex and challenging. How can we best route vehicles such as trucks, school buses, and general passenger buses from one location to another? The problems encountered in answering such questions have an underlying combinatorial structure. For example, either we dispatch a vehicle or we do not, or we use one particular route or another.

This section deals with node-routing problems. Node routing – in contrast to arc routing – refers to routing problems where the key service activity occurs at the nodes (customers), and arcs are of interest only as elements of paths that connect the nodes (Assad and Golden, 1995). We discuss two specific problems that have become prominent in network analysis: the traveling-salesman problem and the vehicle-routing problem. The survey of basic network algorithms is completed by discussing a dynamic programming approach for solving the shortest-path problem with time windows, a problem that occurs as a subproblem in many time-constrained routing and scheduling issues.

At the outset of this section we should note that vehicle-routing algorithms can be applied in one of two modes: first, variable routing and, second, fixed routing. In a variable-routing context, an algorithm is utilized with actual customer delivery requirements to develop routes for the next planning horizon. Fixed routing is applied when customer demands are sufficiently stable to allow use of the same routes repeatedly (Fisher, 1995).

#### 3.1. *The traveling-salesman problem*

The simplest node-routing problem is the traveling-salesman problem. The traveling-salesman problem is a classical combinatorial optimization problem that is simple to state but very difficult to solve. The problem is to find the least-cost tour through a set of nodes so that each node is visited exactly once. The tour starts and ends from a specific location, called the depot. The problem is in a precise mathematical sense difficult, namely NP-complete (“non-deterministic polynomial time”-complete), and cannot be solved exactly in polynomial time. Although there are many ways to formally state the traveling-salesman problem, a convenient way in doing so is an integer linear-programming formulation. Assume a directly

connected network  $G = (N, A)$ , where  $N$  is the node set with  $|N| = N$  and  $A$  is the arc set defined as the Cartesian product of  $N$  with itself (i.e.  $A = N \times N$ ), then the traveling-salesman problem can be formulated as

$$\min_{\{x_{ij}\}} \sum_{i=1}^N \sum_{j=1}^N c_{ij} x_{ij}, \quad (1)$$

subject to

$$\sum_{j=1}^N x_{ij} = 1, \quad \text{for } i = 1, \dots, N, \quad (2)$$

$$\sum_{i=1}^N x_{ij} = 1, \quad \text{for } j = 1, \dots, N, \quad (3)$$

$$(x_{ij}) \in X, \quad (4)$$

$$x_{ij} \in \{0, 1\}, \quad \text{for } i, j = 1, \dots, N, \quad (5)$$

where  $c_{ij}$  is the length of the arc from node  $i$  to node  $j$ , and the  $x_{ij}$  are the decision variables:  $x_{ij}$  is set to 1 when arc  $(i, j)$  is included in the tour, and 0 otherwise.  $(x_{ij}) \in X$  denotes the set of subtour-breaking constraints that restrict the feasible solutions to those consisting of a single tour. The subtour-breaking constraints can be formulated in different ways. But one very intuitive formulation is

$$\sum_{i, j \in S_A} x_{ijk} \leq |S_A| - 1, \quad S_A \subseteq A, 2 \leq |S_A| \leq N - 2, \quad (6)$$

where  $S_A$  is some subset of  $A$  and  $|S_A|$  is the cardinality of  $S_A$ . These constraints prohibit subtours, i.e. tours on subsets with less than  $N$  nodes. If there were such a subtour on some subset of  $S_A$  of nodes, this subtour would contain  $|S_A|$  arcs. Consequently, the left-hand side of the inequality would be equal to  $|S_A|$ , which is greater than  $|S_A| - 1$ , and the constraint would be violated for this particular subset. Without expression (6), the traveling-salesman problem reduces to an assignment problem (Potvin, 1993).

The traveling-salesman problem is a classically hard problem to solve optimally. Enumerating the possibilities works well for small  $N$ . But there are  $(N - 1)!$  candidate itineraries from which the single optimal one must be found. If  $N = 100$ , then the number of possible tours is  $10^{200}$ . Many heuristic algorithms have been devised to solve the problem (for an overview see Lawler et al., 1985; Laporte, 1992). These heuristics were designed to work quickly and to come close to the optimal solution. But they do not guarantee that the optimum will be found. Two broad classes of traveling-salesman heuristics can be distinguished: classical heuristic algorithms and optimization-based algorithms.

Classical traveling-salesman heuristics include tour construction procedures, tour improvement procedures, and composite procedures that are based on both types of techniques. The best known tour construction heuristics gradually build up a tour by selecting each node in turn and inserting them one by one into the current tour. Various metrics may be utilized for the choice of the next node, such as proximity to the current tour. Among the tour improvement procedures, the  $r$ -opt exchange heuristics are the most widely used especially the 2-opt, 3-opt (see Lin, 1965), and the interchange heuristic of Lin and Kernighan (1973). These traveling-salesman heuristics locally modify the current solution by replacing  $r$  arcs in the tour by  $r$  new arcs so as to generate a new improved tour. Characteristically, the exchange heuristics are applied iteratively until a local optimum is found, i.e. a tour that cannot be improved further via the exchange heuristic under consideration. To overcome the limitations associated with local optimality, new heuristics such as tabu search, simulated annealing, and computational intelligence-based techniques may be utilized to escape from local minima (Kirkpatrick et al., 1983; Glover, 1989; 1990, Potvin, 1993). Composite procedures make use of both tour construction and improvement techniques. They belong to the most powerful heuristics used to solve the traveling-salesman problem. The iterated Lin-Kernighan heuristic, for example, can routinely find solutions within 1% of the optimum for traveling-salesman problems with up to 10 000 nodes (Johnson, 1990).

Optimization-based heuristics are very different in character from the classical traveling-salesman heuristics. They apply some optimization algorithm and simply terminate prior to optimality. The most popular technique is branch-and-bound, originally applied to the traveling-salesman problem by Dantzig et al. (1954) and continually refined over the years. Branch-and-bound is a directed enumeration procedure that partitions the solution space into increasingly smaller subsets in an attempt to identify the subset that contains a near-optimal solution. For each subset a bound is calculated that estimates the best possible solution in the subset. The assignment problem relaxation, for example, may be used to generate lower bounds on the optimum. If the bound for a subset indicates that it cannot contain a near-optimal solution, the partitioning process is continued with another subset. The algorithm terminates when there are no subsets remaining. It is worth noting that traveling-salesman problems with a few hundred nodes can be routinely solved to optimality.

### *3.2. The vehicle-routing problem*

There are many ways of generalizing the traveling-salesman problem to match real-world situations. Often there is more than one vehicle, and in these situations the division of stops between variables is an important decision variable (Longley

et al., 2001). In this section we consider the vehicle-routing problem. The problem is to route a fixed number of vehicles through a number of demand locations such that the total cost of travel is minimized and vehicle capacity constraints are not violated. Characteristically, there is a designated location known as the depot where all vehicles have to start and end their tours. There are numerous variations of this basic vehicle-routing problem, including time windows for delivery, stochastic demand, multiple depots with each vehicle in the fleet assigned to a particular depot etc. The problem and its extensions are of substantial practical importance and have resulted in the development of a great many heuristic algorithms including computational intelligence procedures over the past 35 years (for reviews, see Bodin et al., 1983; Golden and Assad, 1986; Fischer, 1995).

We view the vehicle-routing problem as consisting of two interlinked problems: first, finding an optimal assignment of customer orders to vehicles, and, second, ascertaining which route each vehicle will follow in servicing its assigned demand in order to minimize total delivery cost. To provide a precise statement of the problem we introduce notation first and then draw on Fisher and Jaikumar (1981) to specify the vehicle-routing problem as a non-linear generalized assignment problem.

Let  $\mathbf{N} = \{1, \dots, N\}$  be the set of demand locations (customers) and  $\mathbf{K} = \{1, \dots, K\}$  be the set of available vehicles to be routed and scheduled. Consider the network  $G = (\mathbf{V}, \mathbf{A})$ , where  $\mathbf{V} = \mathbf{N} \cup \{0\}$  is the set of nodes with 0 representing the depot of the vehicles, and  $\mathbf{A} = \mathbf{V} \times \mathbf{V}$  is the arc set that contains all arcs  $(i, j)$  with  $i, j \in \mathbf{V}$ .  $c_{ij}$  denotes the cost of direct travel from point  $i$  to point  $j$ ,  $b_k$  the capacity (e.g. weight or volume) of vehicle  $k$ , and  $a_i$  the size of the order of customer  $i \in \mathbf{N}$ , measured in the same units as the vehicle capacity. Define the flow variable  $y_{ik}$  as 0–1 variable equal to 1 if the order from customer  $i$  is delivered by vehicle  $k$ , and 0 otherwise, and  $y_k = (y_{0k}, \dots, y_{nk})$ . Then the vehicle-routing problem can be formulated as the following non-linear generalized assignment problem (Fisher, 1995):

$$\min_{(y_k)} \sum_{k=1}^K f(y_k), \quad (7)$$

subject to

$$\sum_{i=1}^N a_i y_{ik} \leq b_k, \quad \text{for } k = 1, \dots, K, \quad (8)$$

$$\sum_{k=1}^K y_{ik} = \begin{cases} K, & i = 0, \\ 1, & i = 1, \dots, N, \end{cases} \quad (9)$$

$$y_{ik} \in \{0, 1\}, \quad \text{for } i = 0, 1, \dots, N; k = 1, \dots, K. \quad (10)$$

Expressions (8)–(10) are the constraints of the assignment problem and guarantee that each route begins and ends at the depot ( $i = 0$ ), that each customer ( $i = 1, \dots, N$ ) is serviced by some vehicle, and that the load assigned to the vehicle is within its capacity.  $f(y_k)$  represents the cost of an optimal traveling-salesman problem tour of all the points in  $V(y_k) = \{i | y_{ik} = 1\}$  that must be visited by each vehicle  $k$  to service its assigned customers. Defining  $x_{ijk} = 1$  if vehicle  $k$  travels directly from  $i$  to  $j$ , and  $x_{ijk} = 0$  otherwise, the function  $f(y_k)$  can be defined mathematically as

$$f(y_k) = \min \sum_{i=1}^N \sum_{j=1}^N c_{ij} x_{ijk}, \quad (11)$$

such that

$$\sum_{i=1}^N x_{ijk} = y_{jk}, \quad \text{for } j = 0, \dots, K, \quad (12)$$

$$\sum_{j=1}^N x_{ijk} = y_{ik}, \quad \text{for } i = 0, \dots, K, \quad (13)$$

$$\sum_{(i,j) \in \{0,1\}} x_{ijk} \leq |S| - 1, \quad S \subseteq V(y_k), 2 \leq |S| - N. \quad (14)$$

$$x_{ijk} \in \{0, 1\}, \quad \text{for } i = 0, \dots, N; j = 1, \dots, N. \quad (15)$$

The Fisher and Jaikumar (1981) method is the best-known heuristic to solve some mathematical programming approximation of the vehicle-routing problem to optimality. The heuristic replaces  $f(y_k)$  with a linear approximation  $\sum_i d_{ik} y_{ik}$  and solves the resulting linear generalized assignment problem to get an assignment of customers to vehicles (Fischer, 1995). Once this assignment has been made, a complete solution is obtained by applying any traveling-salesman heuristic to get the delivery sequence for the customers assigned to each vehicle.

To obtain the linear approximation, Fisher and Jaikumar (1981) first specify  $K$  “seed” customers  $i_1, \dots, i_K$  that are assigned one to each vehicle. Without loss of generality, customers  $i_k$  can be assigned to vehicle  $k$  for  $k = 1, \dots, K$ . Then the coefficient  $d_{ik}$  is set to the cost of inserting customer  $i$  into the route on which vehicle  $k$  travels from the depot directly to customer  $i_k$  and back. Specifically,  $d_{ik} = c_{0i} + c_{iik} - c_{0ik}$ . Clearly, the seed customers define the direction in which each vehicle will travel, and the assignment problem completes the assignment of customers to routes given this general framework (Fisher, 1995). Seeds are generally chosen with the following rule. Choose the first seed  $s_1$  to be a customer farthest away from the depot. If  $k$  seeds have been chosen, choose  $s_{k+1}$  to solve

$$\max_i \min \left\{ c_{i0}, \min_{j=1, \dots, k} c_{is_j} \right\}. \quad (16)$$

The algorithm can be extended to accommodate a number of variations and generalizations of the above vehicle-routing problem such as the vehicle-routing problem with time windows. This problem consists of designing a set of minimum-cost routes, starting at and returning to a central depot for a fleet of vehicles that services a set of customers with known demands. The service at a customer has to begin within the time window defined by the earliest time and the latest time when the customer allows the start of the service. Time windows can be hard or soft. In the soft time windows case the time window case can be violated at a cost. In contrast, in the hard time window case a vehicle is not permitted to arrive at a node after the latest time to begin service. But, if a vehicle arrives too early at a node, it is allowed to wait until the node is ready for service to begin. The costs involved in time-constrained routing and scheduling consist of fixed-vehicle utilization costs and variable routing and scheduling costs. The latter include distance and travel time costs, waiting time costs, and loading/unloading time costs (Desrosiers et al., 1995).

### 3.3. Constrained shortest-path problems

The section will be concluded by considering the shortest-path (or least-cost) problem with time windows. This problem appears as a subproblem in many time-constrained routing and scheduling problems and, thus, deserves some specific attention. The problem consists of finding the least-cost route between any two specified nodes in a network whose nodes can only be visited within a specified time interval. The description of the problem that follows is based on Desrosiers et al. (1995).

Let  $G = (\mathbf{V}, \mathbf{A})$  be a network where  $\mathbf{A}$  is the set of arcs and  $\mathbf{V}$  the set of nodes  $\mathbf{N} \cup \{o, d\}$ .  $\mathbf{N}$  consists of nodes that can be visited from an origin  $o$  to a destination  $d$ . With each node  $i \in \mathbf{V}$  a time window  $[g_i, h_i]$  is associated. A path in  $G$  is defined as a sequence of nodes  $i_0, i_1, \dots, i_K$  such that each arc  $(i_{k-1}, i_k)$  belongs to  $\mathbf{A}$ . All paths start at time  $g_o$  from node  $i_0$  and finish at  $i_K = d$  no later than  $h_d$ . A path is elementary if it contains no cycles. Each arc  $(i, j) \in \mathbf{A}$  has a positive or negative cost  $c_{ij}$  and a positive duration  $t_{ij}$ . Service time at node  $i$  is included in  $t_{ij}$  for all  $i \in \mathbf{N}$ . An arc  $(i, j)$  in the set  $\mathbf{A}$  is defined to be feasible only if it respects the condition:  $g_i + t_{ij} \leq h_j$ .

The mathematical programming formulation of the shortest-path problem with time windows involves two types of variable: flow variables  $x_{ij}$  with  $(i, j) \in \mathbf{A}$  and time variables  $t_i$  with  $i \in \mathbf{V}$ . Using this notation the shortest-path problem with time windows may be formulated as follows:

$$\min \sum_{(i,j) \in A} c_{ij} x_{ij}, \quad (17)$$

subject to

$$\sum_{j \in V} x_{ij} - \sum_{j \in V} x_{ji} = \begin{cases} +1, & i = o, \\ 0, & \text{for } i \in N, \\ -1, & i = d. \end{cases} \quad (18)$$

$$x_{ij} \geq 0, \quad \text{for } (i, j) \in A, \quad (19)$$

$$x_{ij}(t_i + t_{ij} - t_j) \leq 0, \quad \text{for } (i, j) \in A, \quad (20)$$

$$g_i \leq t_i \leq h_i, \quad \text{for } i \in V. \quad (21)$$

The objective function (17) attempts to minimize the total travel cost. Constraints (18) and (19) define the flow conditions on the network, while time windows appear in constraint (20). Compatibility requirements between flow and time variables are given in expression (21). This non-linear problem with time windows is appealing because it can be shown that if the problem is feasible, then there is an optimal integer solution (Desrosiers et al., 1995).

The problem can be solved by dynamic programming. For the introduction of this approach, define  $Q(S, i, t)$  as the minimum cost of the path routing from node  $o$  to node  $i$  ( $i \in N \cup \{d\}$ ) visiting all nodes in the set  $S \subseteq N \cup \{d\}$  only once, and servicing node  $i$  at time  $t$  or later. The cost  $Q(S, i, t)$  can be calculated by solving the following recurrence equations:

$$Q(\emptyset, o, g_o) = 0, \quad (22)$$

$$\begin{aligned} Q(S, j, t) = \min \{ & Q(S - \{j\}, i, t') + c_{ij} \\ & \text{with } i \in S - \{j\}, t' \leq t - t_{ij}, g_j \leq t' \leq h_i \} \\ & \text{for all } S \subseteq N \cup \{d\} \text{ for } j \in S \text{ and } g_j \leq t \leq h_j. \end{aligned} \quad (23)$$

The optimal solution is given by

$$\min_{S \subseteq N \cup \{d\}} \min_{g_d \leq t \leq h_d} Q(S, d, t). \quad (24)$$

It is worthwhile to note that expression (23) is valid only if  $g_j \leq t \leq h_j$ . If  $t > h_j$ , then  $Q(S, j, t) = Q(S, j, g_j)$ , and if  $t < g_j$ , then  $Q(S, j, t) = \infty$ . The shortest-path problem with time windows is NP-hard in the strong sense. Therefore, the dynamic programming algorithm suggested by Desrosiers et al. (1995) has an exponential complexity, and no pseudo-polynomial algorithm is known for this problem.

#### 4. Concluding remarks

GIS-T, once the sole domain of public sector planning and transportation agencies, is increasingly being used in the private sector to support logistics in general and distribution and production logistics in particular. The cost of the technology is now within the reach of even smaller enterprises. The cost of acquiring the data to populate a GIS for transportation-related applications is falling rapidly. The availability of data is paralleled by GPS services to reference locations accurately. These trends suggest that GIS-T has arrived as a core technology for transportation (Sutton and Gillingwater, 1997).

The performance of GIS-T software largely depends on how well nodes and links and transportation-related characteristics are arranged into a data structure. The data structure must not only represent the complexities of transport networks in sufficient detail but also allow for rapid computation of a wide variety of sophisticated network procedures, such as traveling-salesman problem, vehicle-routing problem, and shortest-path problem algorithms, based on actual network drive time, and not straight-line distances.

#### References

- Assad, A.A. and B.L. Golden (1995) "Arc routing methods and applications," in: M.O. Ball, T.L. Magnanti, C.L. Monma and G.L. Nemhauser, eds, *Handbooks in operations research and management science*, Vol. 8. Amsterdam: Elsevier.
- Atzeni, P., S. Ceri, S. Paraboschi and R. Terlone (1999) *Database systems*. Reading: McGraw Hill.
- Bell, M. and Y. Iida (1997) *Transportation network analysis*. Chichester: Wiley.
- Bespalko, S.J., J.C. Sutton, M. Wyman, J.A. van der Veer and A.D. Sindt (1998) "Linear referencing systems and three dimensional GIS," *1998 Annual Meeting of the Transportation Research Board*, Paper 981404. Washington, DC.
- Bodin, L., B.L. Golden, A. Assad and M. Ball (1983) "Routing and scheduling of vehicles and crews: the state of the art," *Computers and Operations Research*, 10:63–211.
- Dane, C. and C. Rizos (1998) *Positioning systems in intelligent transportation systems*. Boston: Artech.
- Dantzig, G.B., D.R. Fulkerson and S.M. Johnson (1954) "Solution of a large-scale traveling salesman problem," *Operations Research*, 7:58–66.
- Desrosiers, J., Y. Dumas, M.M. Solomon and F. Soumis (1995) "Time constrained routing and scheduling," in: M.O. Ball, T.L. Magnanti, C.L. Monma and G.L. Nemhauser, eds, *Handbooks in operations research and management science*, Vol. 8. Amsterdam: Elsevier.
- Fisher, M. (1995) "Vehicle routing," in: M.O. Ball, T.L. Magnanti, C.L. Monma and G.L. Nemhauser, eds, *Handbooks in operations research and management science*, Vol. 8. Amsterdam: Elsevier.
- Fisher, M. and R. Jaikumar (1981) "A generalized assignment heuristic for vehicle routing," *Networks*, 11:109–124.
- Fohl, P., K.M. Curtin, M.F. Goodchild and R.L. Church (1996) "A non-planar, lane-based navigable data model for ITS," in: M.J. Kraak and M. Molenaar, eds, *Proceedings of the 7th International Symposium on Spatial Data Handling*. Delft.
- Glover, F. (1989) "Tabu search, part I," *ORSA Journal on Computing*, 1:190–206.
- Glover, F. (1990) "Tabu search, part II," *ORSA Journal on Computing*, 2:4–32.
- Golden, G.B. and A.A. Assad (1986) "Perspectives on vehicle routing. Exciting new developments," *Operations Research*, 14:803–810.

- Goodchild, M.F. (1998) "Geographic information systems and disaggregate transportation modeling," *Geographical Systems*, 5:19–44.
- Johnson, D.S. (1990) "Local optimization and the traveling salesman problem," in: G. Goos and J. Hartmanis, eds, *Automata, languages and programming*. Berlin: Springer-Verlag.
- Kirkpatrick, S., C.D. Gelatt and M.P. Vecchi (1983) "Optimization by simulated annealing," *Science*, 220:671–680.
- Kwan, M.-P., R.G. Golledge and J.M. Speigle (1996) "A review of object-oriented approaches in geographic information systems for transportation modeling," Draft document. Santa Barbara: Department of Geography, University of California.
- Laporte, G. (1992) "The travelling salesman problem: an overview of exact and approximate algorithms," *European Journal of Operational Research*, 59:231–247.
- Lawler, E.L., J.K. Lenstra, A.H.G. Rinnoy Kan and D.B. Shmoys (1985) *The traveling salesman problem: a guided tour of combinatorial optimization*. Chichester: Wiley.
- Lin, S. (1965) "Computer solutions of the travelling salesman problem," *Bell System Technical Journal*, 44:2245–2269.
- Lin, S. and B. Kernighan (1973) "An effective heuristic algorithm for the traveling salesman problem," *Operations Research*, 21:498–516.
- Longley, P.A., M.F. Goodchild, D.J. Maguire and D.W. Rhind (2001) *Geographic information systems and science*. Chichester: Wiley.
- Miller, H.J. and S.-L. Shaw (2001) *Geographic information systems for transportation. Principles and applications*. Oxford: Oxford University Press.
- Nyerges, T.L. (1990) "Locational referencing and highway segmentation in a geographic information system," *ITC Journal*, 60: 27–31.
- Potvin, J.-Y. (1993) "The travelling salesman problem: a neural network perspective," *ORSA Journal on Computing*, 5:328–348.
- Spear, B.D. and T.R. Lakshmanan (1998) "The role of GIS in transportation planning and analysis," *Geographical Systems*, 5:45–58.
- Sutton, J. (1997) "Data attribution and network representation issues in GIS and transportation," *Transportation Planning and Technology*, 21:25–44.
- Sutton, J. and D. Gillingwater (1997) "Geographic information systems and transportation – overview," *Transportation Planning and Technology*, 21:1–4.
- Vonderhe, A. and T. Hepworth (1996) *A methodology for design of a linear referencing system for surface transportation*, Project AT-4567 Research Report. Albuquerque: Sandia National Laboratory.

***Part 6***

**GPS APPLICATIONS**

## DEFINING GPS AND ITS CAPABILITIES

JEAN WOLF

*GeoStats, Atlanta, GA*

### 1. Introduction

Determining one's location accurately has always been a fundamental challenge for navigation. Within the world of transport, which focuses on the efficient movement of people and goods through space and time, location determination and navigational guidance are essential. Recently, centuries-old methods of navigation that relied on paper maps, compasses, and stars have been replaced with new technologies based on a man-made satellite-based radionavigation system designed specifically to provide highly accurate, reliable, continuous 24 h, worldwide coverage for location determination. This system was developed by the US Department of Defense, and was originally introduced as the Navigation System with Time And Ranging Global Positioning System (NAVSTAR GPS), and is now referred to simply as GPS.

Two key features of GPS are its all-weather functionality and worldwide coverage area, which allows any user with an inexpensive, widely available GPS receiver to obtain accurate second-by-second three-dimensional position, velocity, and time (PVT) information. Consequently, applications of GPS have proliferated across all areas of transport, beginning with traditional transport areas such as field surveying for road and bridge projects; the collection of road characteristics for geographic information system (GIS) development, and transport agency asset inventory and maintenance applications; and moving rapidly into newer areas such as automatic vehicle location and navigation (AVLN) applications; travel time and delay studies; enhanced mobile source emissions modeling; and travel behavior surveys. Given continued improvements in user technology, including smaller GPS components and lower power demands, it is obvious that GPS capabilities will soon extend beyond vehicle applications into many personal (i.e. wearable) applications.

This chapter introduces the various components and specifications of GPS, but does not attempt to cover everything about this technology. There are many books and journals available that are devoted to explaining the exact science of GPS and

other global navigation satellite systems (GNSS), and these are recommended for the reader interested in more information. Furthermore, even though there are other satellite-based navigation systems, including the Russian Global Navigation Satellite System (GLONASS) and the European-proposed Galileo system, this chapter will concentrate more specifically on NAVSTAR GPS. Finally, although GPS was designed initially for military users and applications, this chapter will focus on civilian features and capabilities of GPS.

## 2. The Global Positioning System

GPS, known originally as NAVSTAR GPS, is a satellite-based radionavigation system designed and developed by the US Department of Defense as a navigational aid. Planning for the system began in the 1960s, with GPS reaching full operational capacity in 1995. This system allows an unlimited number of GPS receivers located anywhere on the earth's surface and in view (or in line-of-sight) of the GPS satellites to accurately determine position, velocity, and time. Although originally developed for military purposes, civilian uses of GPS can be found everywhere. In fact, GPS user technology has been rapidly adapted and applied to a wide range of civilian land, sea, and air applications.

### 2.1. *Overview of GPS*

The GPS system architecture consists of three segments: the space segment, with 24 satellites providing worldwide coverage; the operational control segment, which monitors and controls the space segment; and the user segment. The space segment contains a nominal 24 satellites that orbit the earth every 12 h and are non-symmetrically distributed in six orbital planes of four satellites each – these orbits are nearly circular and equally spaced about the equator at a 60° separation and with an inclination of nominally 55° relative to the equator (Kaplan, 1996). This coverage pattern ensures at least six satellites are always in view at any point on the globe's surface (Figure 1). As new GPS satellites are built with improved functionality, they are deployed on a “launch on need” basis to supplement the existing space segment in which older satellites have outlived their forecasted life-span yet have lost some level of redundancy in one or more critical components (Morris, 2003). The full GPS satellite constellation included 28 satellites as of April 1, 2003, with the most recent satellites deployed where needed for redundancy and eventual replacement within the six orbits.

The operational control segment consists of a master control station located in Colorado Springs, Colorado, in the USA, and five monitor and four uplink (or ground antenna) stations that are geographically distributed around the world.

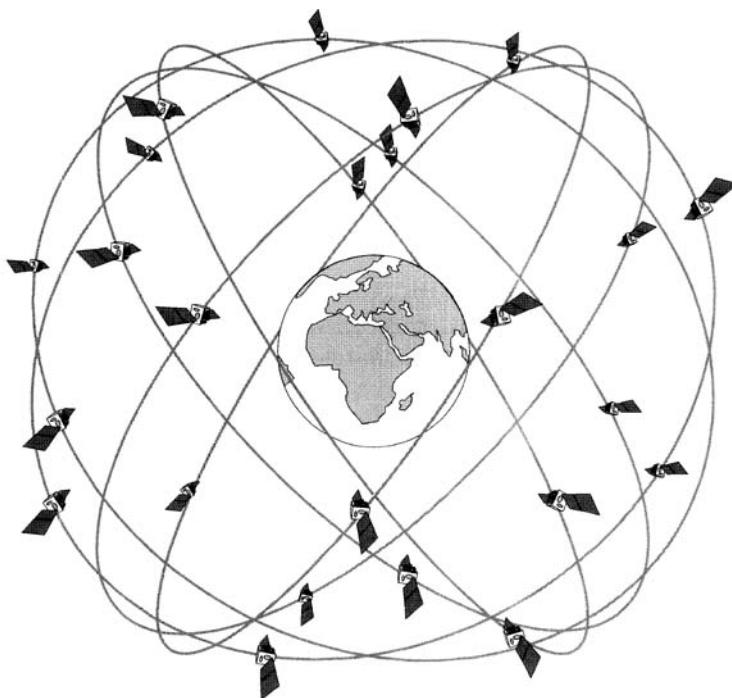


Figure 1. The NAVSTAR GPS operational constellation. (Redrawn from US Department of Transportation, 2000.)

This control segment accurately tracks the GPS satellites (keeping them in their proper orbital positions or slots); updates each satellite's ephemeris, almanac, and clock as needed; and monitors the health and status of each satellite. Figure 2 shows all three segments and their relationship with respect to data flow.

The GPS almanac gives approximate orbital parameters for all satellites, and each satellite broadcasts almanac data for all satellites. Almanac data are not precise, and are, therefore, only considered valid for a certain time period, which has been defined to range from 3 days to several months (Kaplan, 1996; US Department of Defense, 1996). The almanac is used by GPS receivers to predict approximate satellite positions and thereby reduce the search window and associated acquisition time (also referred to as time to first fix). Ephemeris data contain very precise orbital and clock corrections for each satellite; each satellite broadcasts only its own ephemeris. Once received, ephemeris data are considered valid for 3–4 h (Kaplan, 1996).

The user segment receives the data contained in the satellites signals, and uses the data to compute position, velocity, and time. Users of GPS today include

both the military and civilians who require positional information. Military applications include target acquisition, missile guidance, sensor emplacement, coordinate bombing, and remotely piloted vehicle operations. Examples of civil applications include *en route* navigation, flight guidance, fleet management, search and rescue, recreation, theft deterrence, surveillance, and mapping.

## 2.2. PVT determination

GPS computes ground position by using one-way time of arrival ranging. Based on highly accurate atomic frequency standards for satellite clock time, the signal travel times from satellites in view and a ground-based receiver are calculated. Since radio signals travel at the speed of light, these travel times can be used to calculate the distances from the receiver to the satellites in view. The position

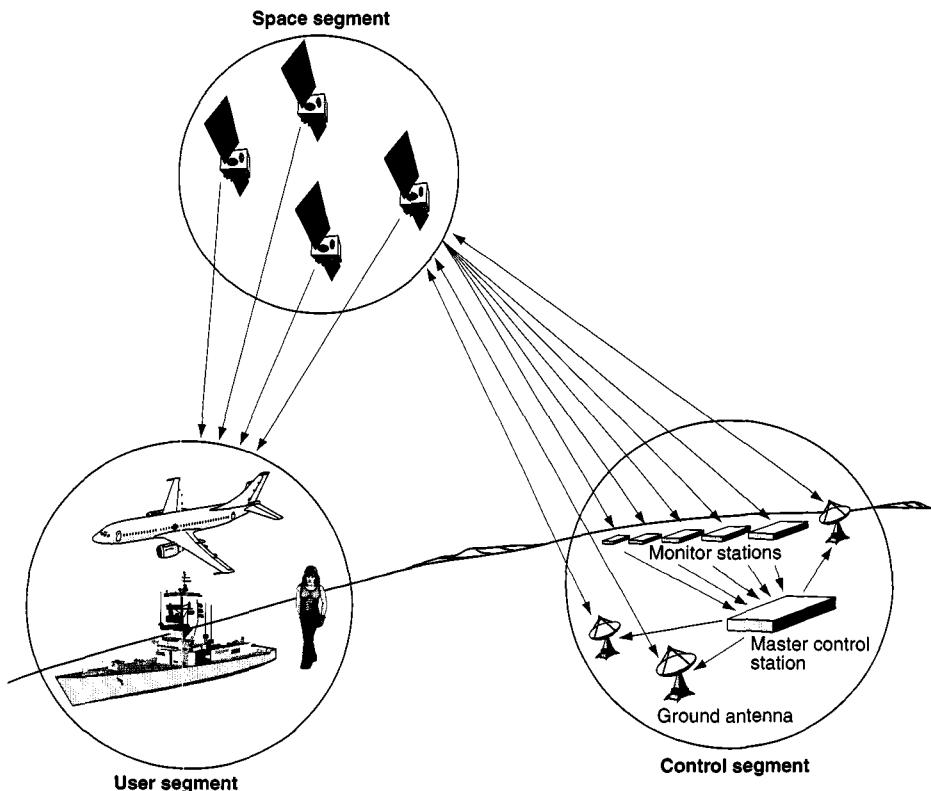


Figure 2. NAVSTAR GPS segments. (Redrawn from US Department of Defense, 1996.)

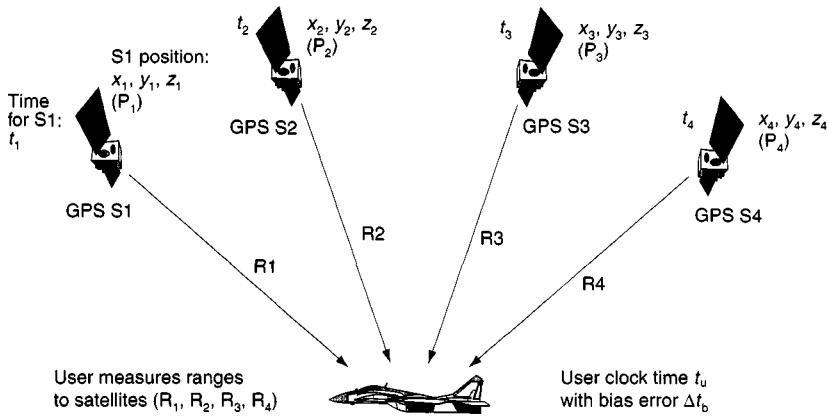


Figure 3. GPS position determination technique. (Redrawn by permission of Navtech Seminars and GPS Supply from McDonald, 1999.)

of the ground receiver's antenna is calculated using trilateration to solve four unknowns: the  $x$ ,  $y$ , and  $z$  coordinates, and the difference between the satellites' clocks and the receiver's internal clock. Figure 3 shows the variables used in GPS-based position determination using four satellites.

The calculations involved for solving the user's position coordinates ( $x_u, y_u, z_u$ ) and clock bias ( $\Delta t_b$ ) are contained in the following four equations:

$$(x_1 - x_u)^2 + (y_1 - y_u)^2 + (z_1 - z_u)^2 = (\tilde{R}_1 - c \Delta t_b)^2,$$

$$(x_2 - x_u)^2 + (y_2 - y_u)^2 + (z_2 - z_u)^2 = (\tilde{R}_2 - c \Delta t_b)^2,$$

$$(x_3 - x_u)^2 + (y_3 - y_u)^2 + (z_3 - z_u)^2 = (\tilde{R}_3 - c \Delta t_b)^2,$$

$$(x_4 - x_u)^2 + (y_4 - y_u)^2 + (z_4 - z_u)^2 = (\tilde{R}_4 - c \Delta t_b)^2.$$

Using a GPS receiver, the user measures four pseudo-range values:

$$R_i = (\tilde{R}_i - c \Delta t_b),$$

where  $R_i$  is the true range,  $\tilde{R}_i$  is the pseudo-range,  $c \Delta t_b$  represents the clock bias range error,  $\Delta t_b$  represents the user clock bias or timing error, and  $c$  is the speed of light.

An alternative method for visualizing the position calculation process is to consider the sphere around each satellite that describes the calculated distance from the satellite to the receiver. The receiver must lie somewhere on the surface of this sphere. Three satellites in view provide three spheres. The intersection of these three spheres yields two points, one on the surface of the earth and another one in outer space, which is automatically discarded. To determine and eliminate

any clock drift, four satellites must be in view to compute a three-dimensional position. A two-dimensional ( $x, y$ ) position can be calculated by using the third satellite to resolve the clock differential rather than for determining the  $z$  coordinate.

There are several methods for determining user velocity. The first method estimates velocity by forming an approximate derivative of the user position, which is acceptable when velocity is relatively constant over the selected time period (Kaplan, 1996). A more common method for user velocity determination is based on the Doppler shift frequency of the GPS satellite signal due to relative motion between the satellite and the receiver. In this method, the satellite velocity vector can be calculated using the ephemeris data, the navigation solution components, and the measured Doppler shifts (McDonald, 1999). Many GPS manufacturers are reluctant to share the details of their velocity calculation algorithms, although most claim velocity accuracy levels within 0.1 m/s (= 0.36 kph) at confidence levels of 95% or better.

### 2.3. Other GNSS

GNSS is a generic term referring to a system containing at least one or more global satellite navigation systems. Beyond GPS, which certainly qualifies as a GNSS, is the Russian GLONASS. GLONASS is similar to GPS in that it is based on a 24-satellite constellation, ground monitoring stations, and a user segment. Also similarly to GPS, GLONASS was designed by the Russian Ministry of Defense for military purposes in the 1970s. There are several design differences between GPS and GLONASS, including a Russian constellation of three orbital planes with eight satellites per plane. Perhaps the biggest difference between the two systems, however, has been the lack of operating satellites for GLONASS. According to the official GLONASS website (<http://www.rssi.ru/SFCSIC/english.html>), more than 70 satellites have been launched since 1982, yet only 10 are currently operational. On a positive note, Russia has launched three satellites per year for the last 3 years, including three launched as recently as December 2002, and, at this rate, it is possible that Russia may meet its latest plan to reach an operational 24-satellite constellation by 2007 (*GPS World*, 2003).

The European Commission and the European Space Agency (ESA) have been planning for Europe's own GNSS – Galileo. If Galileo is implemented, it will be the first GNSS designed primarily for civilian use. One primary motivation for the development of a GNSS independent of GPS is that Europeans and civilians worldwide would be equally independent of US government (more specifically, the US Department of Defense) control of GPS. The latest schedule for Galileo calls for a test satellite to be in place by end of 2005. Then, four spacecraft initially, orbiting in two planes, along with the ground segment, will constitute the start of the system. At a total estimated cost of 3.3 billion euros, Galileo is scheduled to

reach full operational capability in 2008. However, the member countries have already caused schedule delays as they compete for development work from the ESA. Given these schedule delays and other program obstacles, it is possible that the full operational capability might not be achieved until 2012 (Divis, 2003a).

#### 2.4. GPS user technologies

There are hundreds of commercially available GPS receivers and antennas on the market today. In *GPS World's* 2003 Annual Receiver Survey, more than 500 receivers produced by more than 70 manufacturers were listed. Similarly, *GPS World's* 2003 Antenna Survey lists 241 antennas from 30 manufacturers. These products have prices ranging from the low hundreds to tens of thousands of dollars; prices typically reflect the level of precision available, where the cheapest receivers produce simple uncorrected position information (5–20 m with selective availability (SA) off), and the most expensive survey-quality receivers can provide sub-centimeter level accuracy using real-time kinematic differential GPS (DGPS, a method discussed later in this chapter). These receivers support a variety of uses (e.g. marine, aircraft, vehicle, and personal) and correction modes (e.g. autonomous, real-time DGPS, post-processing DGPS, and inverse DGPS), come in a range of forms (e.g. OEM boards and chip sets, PCMCIA cards, sensors, handheld computers), and have a wide selection of antennas that also are available in an assortment of forms, features, and configurations. This vast array of options has been developed to meet a maturing GPS marketplace as user groups increase their acceptance of this relatively new positioning technology.

Examples of the variety found in GPS user equipment can be seen in Figure 4. Most often, the exact application of GPS technology will determine the appropriate form factor to house the receiver and antenna, as well as the micro-controller or central processing unit that contains the control logic for storing and downloading GPS data, and the power cord and/or power supply itself. Since 2000, there have been many advancements in other electronic devices that now either support integrated or plug-in GPS technology – these include personal digital assistants (or PDAs such as Palm OS or Pocket PC OS handheld PCs), cell phones, and even watches. Further technology advancements realized in late 2002 that reduce the size and power demand of GPS chip sets offer the promise of GPS integration into any device.

#### 2.5. GPS receiver output

Most GPS receivers output information using a subset of available NMEA 0183 GPS message formats. NMEA 0183 is the National Marine Electronics Association's



Figure 4. An assortment of typical GPS user products. (Reproduced by permission of Garmin International.)

0183 ASCII interface standard for marine electronic devices. This standard was designed to define electrical signal requirements, data transmission protocol and timing, and specific sentence formats for a 4800 baud serial data bus (National Marine Electronics Association, 2002). As defined in the NMEA 0183 statement of scope, "the standard is developed to permit ready and satisfactory data communication between electronic marine instruments, navigation equipment, and communications equipment when interconnected via an appropriate interface." Examples of such equipment include weather instruments, timekeepers, velocity sensors, radar, Loran C, heading sensors, any GNSS (including GPS and GLONASS), and electronic chart systems.

The NMEA 0183 standard calls for data communication in the form of coded "sentences." Between the beginning and end of each sentence are a number of data fields separated by commas. The first field in any sentence (field 0) begins

with the two-letter talker mnemonic code followed by the three-letter sentence ID. The talker ID for the GPS sentences is “GP.” Data then follows in the standard format for that sentence. If any data for a given field is not available, the field is left empty.

For example, the Garmin GPS 35 TracPak combination receiver/antenna (as seen in Figure 5) has a 1 s factory-set default rate for message generation at 4800 baud, but this can be modified if the user wants to vary the baud rate or output sentences (Garmin, 2000). All data are generated in sentence form, and the sentences are transmitted contiguously. This receiver outputs coordinated universal time (UTC, also referred to as Greenwich mean time), the date, and the time of day in its sentences. Prior to the initial position fix, the on-board clock of the receiver provides the date and time; after the first position fix, GPS satellite data are used to determine the date and time.

Four commonly transmitted NMEA 0183 GPS sentences include (National Marine Electronics Association, 2002):

- RMC – recommended minimum specific GNSS data;
- GGA – GPS fix data;
- GSA – GNSS DOP and active satellites;
- GSV – GNSS satellites in view.

Note that if the GNSS is GPS, then each sentence will start with GP – as in GPRMC, GPGGA, GPGSA, and GPGSV. Box 1 shows the RMC sentence structure as specified by the NMEA standards. Some GPS receivers also transmit proprietary messages in addition to NMEA sentences.

Within the context of transport studies, it is obvious that these sentences, when collected on a second-by-second basis, provide highly useful information for creating inventories of transportation infrastructure/assets as well as for evaluating transport system performance and user behaviors. Critical variables made available by GPS include the date and time (synchronized to the US Naval Observatory’s master clock), latitude and longitude, altitude, speed, and heading. Other variables provided give insight on the accuracy of the PVT calculations; these include the positioning system status field and mode indicator, the position dilution of precision (PDOP) and the horizontal dilution of precision (HDOP), and the number of satellites in use. Example GPS data recorded by a GeoStats GeoLogger is shown in Table 1. A few of these variables will be discussed in more detail in the following section on GPS performance measures.

## 2.6. GPS performance measures

GPS provides two services: the Precise Positioning Service (PPS), which is available only to US authorized military users; and the Standard Positioning

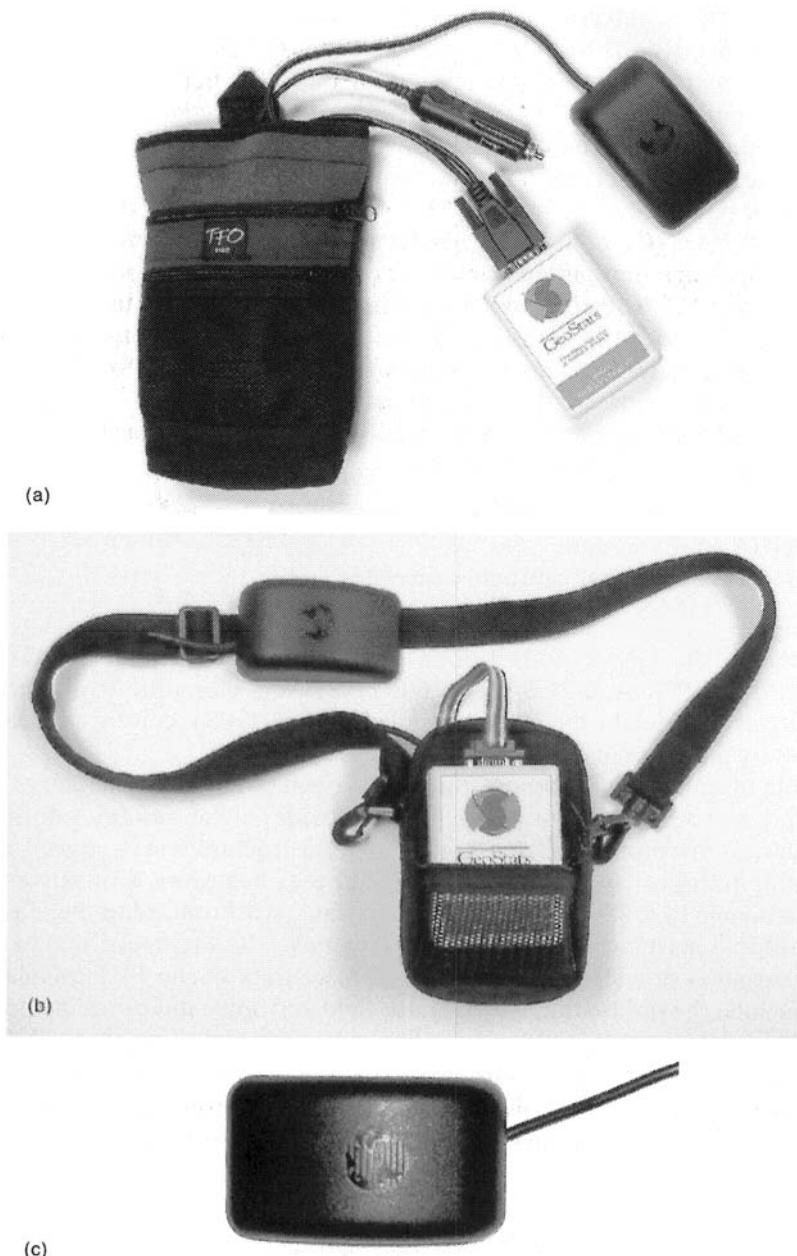


Figure 5. The Garmin GPS 35 TracPak receiver/antenna shown in (a) vehicle-based and (b) person-based logging systems. (c) An enlargement of the receiver.

Box 1

Sentence specifications for recommended minimum specific GNSS data (RMC)

```
$GPRMC,<1>,<2>,<3>,<4>,<5>,<6>,<7>,<8>,<9>,<10>,<11>*hh<CR><LF>
<1> UTC of position fix (hhmmss.ss format)
<2> Status (A = data valid, V = navigation receiver warning)
<3> Latitude (ddmm.mmmm format)
<4> Latitude hemisphere (north or south)
<5> Longitude (ddmm.mmmm format)
<6> Longitude hemisphere (east or west)
<7> Speed over ground (0.0 to 999.9 knots)
<8> Course over ground (0.0–359.9°, true)
<9> UTC date of position fix (ddmmyy format)
<10> Magnetic variation (000.0 –180.0°)
<11> Magnetic variation direction (east or west; westerly variation adds to true course)
```

Table 1  
Example GPS data recorded by a GeoStats GeoLogger

Date (UTC)	Time (UTC)	Latitude	Longitude	Heading	Speed	HDOP	No. of satellites
11/12/2002	19:21:38	33.82680	-84.33285	349	2.6	1.8	6
11/12/2002	19:21:39	33.82682	-84.33287	337	2.7	1.8	6
11/12/2002	19:21:40	33.82682	-84.33287	337	2.7	1.8	6
11/12/2002	19:21:42	33.82682	-84.33292	236	7.7	1.8	6
11/12/2002	19:21:43	33.82680	-84.33295	231	5.8	1.6	7
11/12/2002	19:21:44	33.82678	-84.33297	231	5.8	1.6	7
11/12/2002	19:21:45	33.82667	-84.33313	202	8.4	2.2	6
11/12/2002	19:21:46	33.82663	-84.33315	196	8.8	1.8	6
11/12/2002	19:21:48	33.82658	-84.33315	189	9.5	1.8	6
11/12/2002	19:21:49	33.82655	-84.33315	183	10.4	1.9	6

Service (SPS), which is designed for civilian use. Each service has specifications for performance levels that the US government commits to provide to GPS users. The SPS performance standards were released publicly in 1996 and then updated in 2001 after SA – the US government’s intentional degradation of SPS accuracy levels for military purposes – was discontinued. SPS performance standards are provided for constellation management, service availability, service reliability, accuracy, and status/problem reporting (US Department of Defense, 2001).

#### *Constellation management standard*

The current architecture for the GPS constellation identifies nominal orbit slots for a 24-satellite constellation and defines tolerances for establishing and maintaining

satellites within the slots. The relative spacing between the satellites within a given orbit is maintained such that the position solution geometry criterion (defined in the following service availability standard) is continuously sustained for the core 24-satellite constellation.

#### *Service availability standard*

There are two measures established for service availability – one based on PDOP values and the other based on horizontal and vertical service availabilities. PDOP reflects the dispersion of satellites in the sky in which better distributions receive lower PDOP values and poor dispersions, such as when all in-view satellites are bunched together in a given area of the sky, are given higher PDOP values. Hence, the PDOP service availability standard is defined in terms of the percentage of time that the constellation geometry provides a PDOP value of 6 or less. These standards are set at a minimum acceptable level of 98% globally and 88% for the worst site. To support these standards, 24 operational satellites must be available on orbit with 95% probability over any given day. These standards explain the importance of each satellite staying within the constellation management standard.

The commitment for maintaining acceptable levels of PDOP combined with the service reliability standards led to the definition of the 99% horizontal and vertical service availability standards for an average location as well as the 90% horizontal and vertical service availability standards for a worse-case location. These standards are defined for receivers located in space (known as SIS, or signal in space), which does not include line-of-sight satellite blockages, atmospheric distortion errors, multipath effects (i.e. signals that bounce off of other objects before reaching the receiver, thus producing inaccurate signal travel times), or receiver-specific noise or errors. Both the 99% average and the 90% worse-case standards are set at 36 m horizontal and 77 m vertical at a 95% threshold.

#### *Service reliability standard*

The US government commits to provide SPS service reliability at 30 m or less, measured as an SIS user range error, for a minimum of 99.94% of the time, based on 1 year daily global averages. Defined as such, GPS service reliability specifically addresses how well the system can maintain SIS user range errors within a specified reliability error level.

#### *Accuracy standard*

Commitments for positioning and timing accuracy were established for global averages and worst site positioning; the 95% all-in-view horizontal error (SIS

only) is set at less than or equal to 13 m for the global average, and at less than or equal to 36 m for worst case positioning. For the 95% all-in-view vertical error (SIS only), the global average standard is less than or equal to 22 m, and the worst site positioning standard is less than or equal to 77 m. Finally, the time transfer accuracy standard is defined as less than or equal to 40 ns time transfer error for 95% of the time (SIS only).

### *GPS status and problem-reporting standard*

The US government will issue NANUs (Notice: Advisory to Navigation Users) any time that there are changes or anticipated changes to the GPS constellational status or performance capabilities. As the control segment of GPS monitors the military service (PPS) and not SPS in near real time, there are time delays associated with these civilian notices. The commitment provided is a 48 h notice prior to any scheduled event that will affect GPS service and “as soon as possible after the event” for any unscheduled outages or problems. This performance standard is tied to the concept of GPS integrity, which is the ability of a navigation system to provide timely warnings to users of the system when it should not be used. Accordingly, great emphasis has been placed on GPS integrity monitoring, the result of which has been the implementation of two methods known as receiver autonomous integrity monitoring and the wide-area augmentation system.

## *2.7. Standalone GPS position accuracy and augmentations*

Although the GPS performance standards for accuracy are set based on signal-in-space user range errors, which include satellite clock and ephemeris prediction errors, many other sources of errors are excluded from this definition. Until recently, the largest source of error was SA, which is the intentional introduction of satellite ephemeris errors by the US Department of Defense. SA degraded the precision of GPS accuracy up to 100 m for non-military users. SA was terminated on May 1, 2000, thus improving the accuracy levels of GPS receivers located around the world to, typically, 10–15 meters overnight. Since then, gradual improvements in this accuracy level have been seen, attributed to the launching of advanced-technology satellites, which improve both the accuracy performance and coverage area of the system. The remaining error is attributable primarily to satellite clock and ephemeris prediction errors, atmospheric delay errors (specifically ionospheric and tropospheric errors), multi-path errors, and receiver noise. Table 2 presents a typical GPS error budget at the 1 SD level, broken down by segment and error source (Kaplan, 1996). The total system user equivalent

Table 2  
GPS error budget without selective availability

Segment source	Error source	GPS 1 SD error (m)
Space	Satellite clock stability	3.0
	Satellite perturbations	1.0
	Other (thermal radiation, etc.)	0.5
Control	Ephemeris prediction error	4.2
	Other (thruster performance, etc.)	0.9
User	Ionospheric delay	5.0
	Tropospheric delay	1.5
	Receiver noise and resolution	1.5
	Multi-path	2.5
	Other (interchannel bias, etc.)	0.5
System UERE	Total (RSS)	8.0

range error (UERE) for this budget is 8.0 m, which is calculated as the root sum square (RSS) of the individual error components.

Standalone GPS position determination can be augmented to enhance system accuracy performance. The most common augmentation technique is referred to as DGPS, or differential correction. DGPS is a technique that eliminates both SA-introduced errors and atmospheric errors in either a real-time or post-processing mode, and results in accuracy levels of typically 3–5 m when moving, and sub-meter when surveying at a fixed position. Differential correction is the technique of reducing GPS errors by collecting data with two units simultaneously. A common procedure is to use a base station receiver whose position is precisely known. The satellite data collected at this known position allow for the computation of corrections for the GPS signals. These corrections are then applied to the data collected by another receiver located elsewhere, eliminating atmospheric effects and multi-path errors. Differential corrections can either be sent by the base station to the receiver in real time (real-time DGPS) through a satellite or ground-based radio link, or can be stored on the base station and applied later to the field receiver's data using a process known as post-processing differential correction (post-processing DGPS).

Real-time differential corrections applied to the GPS receiver itself can be provided via a DGPS receiver (built-in or separate) receiving correction signals from space-based radio transmissions, such as those broadcast from OmniSTAR or Thales geostationary satellites and made available for a subscription fee, or ground-based systems such as those broadcast from the US Coast Guard's nationwide DGPS system – a free radio-beacon signal transmitted from “retired”

US Air Force Ground Wave Emergency Network (GWEN) sites that have been converted to NDGPS beacons. Another variation of real-time DGPS is known as inverse differential correction, where uncorrected GPS field data are transferred back to a base station in real time through a wireless communication link; in this scenario, the corrections are applied at the base station. This method offers several advantages with respect to rover data storage and processing for applications that do not require high levels of accuracy in the field. An evaluation of the various real-time DGPS options is available from Wolf and Thittai (2000).

Post-processed differential corrections can be collected at a fixed and known position (e.g. at a base or reference station) and then applied later to the GPS data. Recently, there have been a number of Internet-based sources for regional differential corrections that are collected continuously. One critical requirement for post-processing DGPS is that the GPS data collection must include the full pseudo-range data for all the satellites in view, which requires significantly larger quantities of data to be logged during data collection.

Uncorrected GPS data are referred to as autonomous or raw data. Given the marked improvement in accuracy levels of raw GPS data in today's non-SA world, differential GPS is not as common a requirement for many transport applications. Many off-the-shelf GPS receivers are providing uncorrected accuracy levels consistently in the 5–15 m range. However, for survey-grade or GIS base map data collection work, DGPS is still necessary.

There are also several other technologies that are currently being integrated with GPS receivers for improved capabilities in position determination and/or transmission. For example, inertial navigation systems, also known as dead-reckoning devices, are now used to augment and/or enhance positional data for areas in which GPS and/or differential correction signal reception is poor or non-existent. These areas include those with significant urban canyons and tunnels.

## *2.8. Free satellite-based augmentation systems*

Within the past decade, the USA, Europe, and Japan have each initiated plans for their own free satellite-based augmentation system (SBAS). In the USA, the Federal Aviation Administration has developed the Wide Area Augmentation System (WAAS); in Europe the ESA has begun work on the European Geostationary Navigation Overlay Service (EGNOS), and in Japan, the Japanese Civil Aviation Bureau has started on their Multi-Function Satellite (MTSAT) Satellite Augmentation System (MSAS). As region-based systems, it is critical that these SBASs are compatible to ensure that all systems can be integrated into a seamless worldwide navigation system. Hence, several interoperability working

groups have been created, including EGNOS/MSAS and EGNOS/WAAS. Since WAAS is the first free SBAS to reach its initial operational capacity, a brief description of its features is included here.

WAAS is a US-based system of 25 ground reference stations, two master stations, and at least two geostationary satellites that provide free GPS signal corrections for WAAS-enabled receivers located in the USA. Although intended primarily to increase the reliability, integrity, accuracy, and availability of GPS for aviation users *en route* and through precision approach phases of flight within the USA, ground-based users can also benefit. Many recent models of commercially available GPS receivers are WAAS enabled, which means no special receiving equipment is needed.

WAAS signals can improve GPS position accuracy to within 5 m horizontal and 7.6 m vertical for 95% of the time by correcting for GPS signal errors caused by satellite orbit errors, clock drift, and atmospheric distortion. In addition, WAAS also provides critical integrity information about the health of each GPS satellite. However, the geostationary satellites, with fixed positions over the equator, do not provide full coverage across the USA, especially in areas where views of the southern horizon are blocked.

Although WAAS is a few years behind its original schedule, it reached its initial operational capability in 2003, with 95% availability of signals over at least 75% of the continental USA, and will reach its final operational capability by 2006, with 99.9% availability over 100% of the USA. To achieve these improved levels of availability and coverage, one or more additional geostationary satellites will be added.

## 2.9. *GPS modernization (or GPS III)*

The US government has been planning for a complete modernization of the current GPS. This upgrade is referred to as GPS III. By launching satellites with additional functionality, the USA hopes to expand the current GPS services to include a new military code and several civilian frequencies. GPS III satellite launches are scheduled to begin in 2012, but recent US budget discussions indicate that as much as US \$347 million could be cut from the total GPS III budget through 2007 (Divis, 2003b). Unfortunately, the civilian requirements (led by agencies such as the US Department of Transportation) have neither the power nor the funding behind them to stay on the priority list. However, it is feasible that the GPS III program can be revived within a few years with higher levels of funding and still meet the implementation timeline proposed. At present, though, much remains to be determined with this ambitious and civilian-focused program.

### 3. GPS capabilities for transport

Given the focus of transport on the efficient movement of people and goods through space and time, GPS offers an incredible utility across a broad range of application areas, and, given its young age, it would seem that the civilian world has only started to tap its capabilities. In the following chapters, a selection of GPS applications in the transport arena will be presented. These application areas include studies of household travel, vehicular travel, and traffic operations, as well as centerline mapping, inventory management, automated vehicle location, and safety. There are other references available that focus on GPS applications in transport, including several Transportation Research Board syntheses (Czerniak and Reilly, 1998; Czerniak, 2002) and a US Department of Transportation document (US Department of Transportation, 2000). Consequently, this last section will focus on a few of the latest trends in GPS technology applications for transport.

#### 3.1. Highway, transit, airport, and seaport traffic control and security

With recent world events forcing attention on civilian safety and security, funding priorities have also shifted to areas related to controlling and monitoring the transport infrastructure with respect to security. For example, congestion mitigation previously focused on air quality benefits has now been refocused on transport security needs, e.g. the critical need to keep traffic moving so that vehicles are not parked at locations vulnerable to explosions.

#### 3.2. E911

In the 1990s, the US Federal Communications Commission (FCC) issued an E911 (emergency call) mandate to compel wireless telecommunications carriers to enable the automatic location identification of emergency callers as a result of the rapid increase of cell phone usage. Owing to a range of technology and implementation challenges, the E911 schedule has slipped significantly from its original goal for all cellular phones to convey the caller's position to within 100 m for 67% of all calls and within 300 m for 95% of all calls by October 31, 2001. This network-based accuracy goal itself was expanded to include handset-based solutions, in which the cell phone has GPS capabilities (referred to as assisted GPS, or A-GPS) for location determination. The handset-based solutions must provide the caller's position within 50 m for 67% of the calls and within 150 m for 95% of the calls. Given that more than 10 million cell phones with integrated GPS were sold in the 18 months between September 2001 and March 2003, it seems

that the technology solution will be GPS in handsets, but the emergency dispatch centers (also known as public service access points) will need software and hardware to receive the E911 location data. The latest FCC schedule established a 4 year roll-out beginning October 1, 2001 and to be completed by December 31, 2005.

### *3.3. Location-based services*

Partially as a result of the E911 mandate, the ability to identify the location of a cell phone also offers the ability for targeted marketing toward anyone leaving his or her home with a location-aware wireless device (such as a cell phone or PDA). Location-based services (LBS) are defined as the ability to find the geographic location of a mobile device and to provide services based on this location (Batty, 2003). Although GPS works well outdoors, local positioning system technologies have been developed that may provide good indoor coverage. These technologies include ultra-wideband (UWB), which enables high-bandwidth wireless networks that can supply highly accurate tracking information. A-GPS uses a wireless network, with its own GPS receivers, to predict the GPS signal that a given handset will receive and to relay that information to the handset; this technology may also provide sufficient indoor location information. One goal of consumer LBS is to offer product and service-specific discounts and deals as a consumer approaches a particular vendor's location. It is obvious that if these "wearable tracking" devices were to share their data with traffic management operators or transport researchers, much could be gained in the area of transport system monitoring and travel behavior analysis. The privacy issues associated with such data, however, may prevent these data sets from ever becoming available to public agencies.

### *3.4. Combined measures of travel, physical activity, and health*

As GPS receivers have become more wearable, researchers in the health and environment fields have begun examining combined measures of travel behavior and physical activity. Within the USA, the Robert Woods Johnson Foundation has recently devoted millions of research dollars to gain a better understanding of the relationships between travel modes, the physical or built environment, and the levels of activity experienced by individuals moving through and living within these environments. GPS technology enables the collection of spatial and temporal data in tandem with other wearable activity monitors, as well as spatially accurate inventories of environmental attributes that may impact activity levels. Figure 6 shows several GPS traces of activity spaces, along

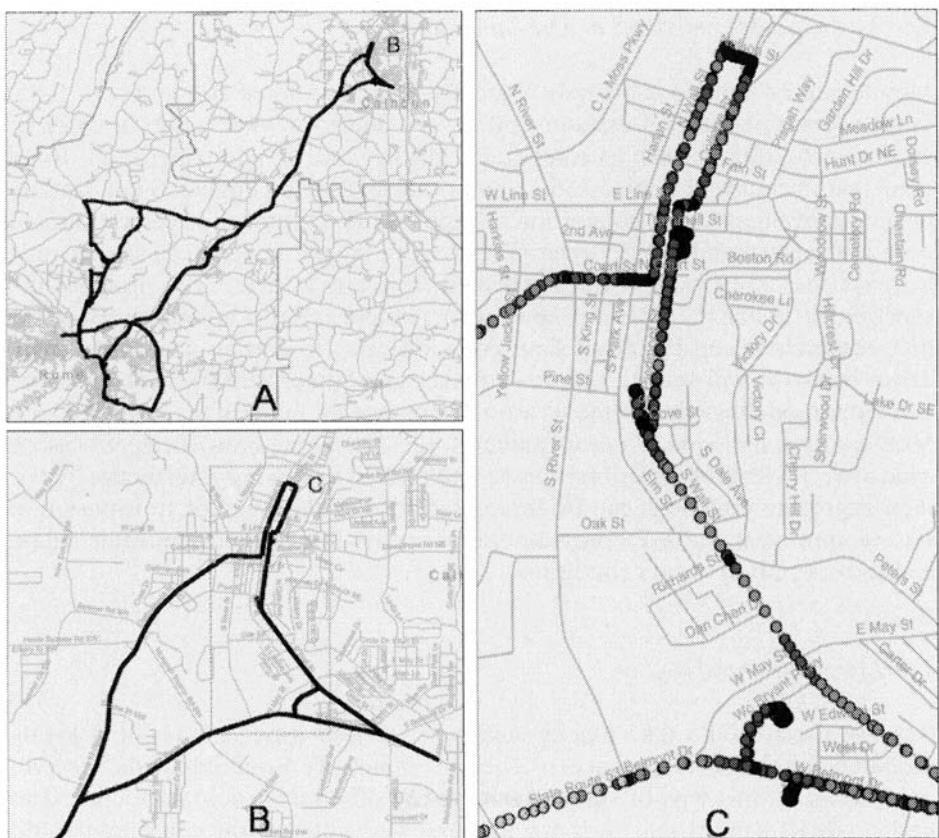


Figure 6. Maps of regional and local activity spaces and levels.

with some activity level data in map C. Although reproduced in grayscale here, the original trace is in color, and the color-coded activity levels reflect energy expenditure as measured by an activity monitor (or accelerometer). The US Environmental Protection Agency is now looking at innovative technologies for remote data collection to support the upcoming US National Children's Study, which is a large long-term study of environmental influences on children's health and development. Technologies under consideration include the use of PDAs, wireless communications, the Internet, GPS, and devices that measure indoor or outdoor air quality, as well as devices that collect and transmit health and biological measures, such as blood pressure, heart rate, and weight.

### *3.5. Mobile source emissions analysis and modeling*

Mobile source emissions analysis is greatly enhanced when coupled with GPS. Vehicle gram-per-second emission rates vary due to a wide range of external operating conditions such as road grade, traffic control, and congestion. Road grade can increase or decrease the amount of load on an engine. Signalization, turning movements, and congestion all impact the amount of acceleration and deceleration demanded by a driver. These operating parameters can significantly increase the base emission rates expected from any vehicle. GPS provides the capability to match the emission data stream to these external influences. By doing this, researchers can better allocate emission rates to certain combinations of driver behavior and vehicle activity (Greaves and Somers, 2003). Mobile source emissions modeling improvements are also possible for many of the same reasons. Mobile source emission inventory models forecast the emissions of vehicles over a wide area. If elevated emission rates can be related to road or intersection types, then aggregate emissions can be forecasted for a wide range of transportation system improvements. GPS provides the necessary linkage between vehicle data and transportation system conditions.

### *3.6. Long-term travel studies*

Because passive GPS data logging enables extended travel studies, it is highly probable that the development of a “Nielsen” family of households or vehicles will soon occur. In this type of study, a sample of households could be recruited to participate for a multi-year period. Each person or vehicle of the household would be instrumented with a GPS data logger and a communications link to transfer the GPS data back to a central location on a regular basis. In a method similar to that used for Nielsen television viewer ratings, the travel and activity patterns of this sample could be analyzed and monitored over time to assess the impact of congestion or transportation control measures on travel behavior (Wolf, 2000).

## **Appendix: Internet resources for GPS**

Four websites that provide a wide range of supporting information about GPS are:

- <http://www.igeb.gov/>  
The website of the Interagency GPS Executive Board.
- <http://gps.losangeles.af.mil/>  
The website of the NAVSTAR Global Positioning System Joint Program Office.

- <http://www.navcen.uscg.gov/>  
The US Coast Guard Navigation Center website.
- <http://gps.faa.gov/>  
The website of the US Federal Aviation Administration Satellite Navigation Product Teams.

## References

- Batty, P. (2003) "Real-time spatial applications drive industry growth," *Geoworld*, 16:30–33.
- Czerniak, R. (2002) *Collecting, processing, and integrating GPS data into GIS. NCHRP synthesis 301*. Washington, DC: National Academy Press.
- Czerniak, R. and J. Reilly (1998) *Applications of GPS for surveying and other positioning needs in departments of transportation. NCHRP synthesis 258*. Washington, DC: National Academy Press.
- Divis, D.A. (2003a) "Washington view: Galileo stuck in political traffic," *GPS World*, 14:10.
- Divis, D.A. (2003b) "Washington view: GPS budget slashed," *GPS World*, 14:10–12.
- Garmin (2000) *Garmin GPS 35 LP TracPak™ technical specification*. Olathe: Garmin.
- GPS World* (2003) "Global view: new birds fly for Glonass, GPS," *GPS World*, 14:14.
- Greaves, S.P. and A. Somers (2003) "Insights on driver behaviour: what can global positioning system (GPS) data tell us?" in: *Proceedings of the 21st ARRB and 11th REAAA Conference*. Cairns.
- Kaplan, E.D. ed. (1996) *Understanding GPS: principles and applications*. Boston: Artech House.
- McDonald, K. (1999) *Fundamentals of GPS I and II. Institute of Navigation GPS-99 tutorials*. Nashville: Navtech Seminars and GPS Supply.
- Morris, J. (2003) "GPS III options to be presented to Teets in mid-April," *Aerospace Daily*, April 1.
- National Marine Electronics Association (2002) *NMEA 0183: Standard for interfacing marine electronic devices. Version 3.01*. Severna Park: NMEA.
- Wolf, J. (2000) "Using GPS data loggers to replace travel diaries in the collection of travel data," dissertation. Atlanta: Georgia Institute of Technology, School of Civil and Environmental Engineering.
- Wolf, J. and R. Thittai (2000) "Real-time differential correction options for GPS route choice data collection, in: *Proceedings of the 79th Annual Meeting of the Transportation Research Board*, Paper 00-1034. Washington, DC.
- US Department of Defense (1996) *NAVSTAR GPS user equipment introduction*. Washington, DC: US Department of Defense.
- US Department of Defense (2001) *Global Positioning System standard positioning service performance standard*. Washington, DC: Assistant Secretary of Defense for Command, Control, Communications, and Intelligence.
- US Department of Transportation (2000) *National civilian GPS services*. Washington, DC: US Department of Transportation.

*Chapter 24*

## GPS, LOCATION, AND HOUSEHOLD TRAVEL

PETER R. STOPHER

*The University of Sydney*

### 1. Introduction

As has been pointed out in earlier chapters, there is a need within transport planning and allied areas to have accurate information on the geography of people's travel. However, most people are not qualified geographers and do not understand or relate to the surrounding geography in a way that permits them to report it in a precise enough manner for the purposes of planning (Stopher et al., 2002a). Most people know their own home address. Some people know their workplace address, but those who do not have to send mail to or from their workplace often do not know it. Most people do not know the street address of the school where their children attend, although they probably know the name of the school and may know the name of the street on which it is located. Almost no one could provide the street address of the grocery stores that they use most frequently. They know the name of the store, and probably know the suburb in which it is located (although this is not always the case), but few would know the postal address. So far as other locations that household members visit, such as medical offices, church, gym, it is unlikely that any would know the street address, except, perhaps, for a friend's or relative's home that is visited infrequently, and for which the address has to be looked up.

The second area in which knowledge of location is desirable is the route a person chooses when traveling by car, bicycle, foot, or other private means of travel. Few people could describe routes taken frequently, such as the drive to work, in terms of the sequence of street names. This becomes even more difficult in a city like Sydney, where street names change along the length of what appears to be a continuous street. The person traveling along that street may know the name of it at the beginning, but is unlikely even to be aware that the name has changed part way along. Routes taken infrequently may be described by street name, if this is the way in which directions are given or prepared. However, many people travel by landmarks, not street names. A route is often described in a similar way to the following: "You leave home and travel straight ahead to the first

roundabout, then turn left to the corner where you will see a large white house with carriage lamps at the entrance, where you turn right. Go straight ahead until you see the supermarket. Then turn left immediately after the car park for the supermarket ..." This is an accurate description of the route, but is unlikely to be of use to the transport planner who is trying to define the route in a systematic manner in a geographic information system (GIS).

For as long as transport planning has been done, this has been the dilemma of the transport planner. On the one hand, accurate data are needed about the locations of the origins and destinations of each trip that has been taken by each household member, and detailed route choice information is desired, in order to be able to improve upon the problematic shortest path assignment algorithms. On the other hand, the transport planner has to consider the burden placed upon the respondent to a household travel survey, who is largely unable to provide addresses and cannot provide routing information that is at all useful. Furthermore, to request detailed routing information on each trip undertaken by each household member would add unacceptably to what is already considered to be a burdensome survey.

## 2. GPS as a solution

As outlined in Chapter 23, the Global Positioning System (GPS) was developed by the US Department of Defense and deployed in the 1980s. By the early 1990s, the public were able to obtain devices that would use GPS to find their location anywhere on the face of the earth. For a number of years, the military used a system called selective availability (SA) to intentionally downgrade the positioning information received by civilian receivers, so that positioning was not more accurate than about  $\pm 100$  m. In the late 1990s, SA was removed, and the positioning accuracy, without the use of any type of differential correction, moved to being within a few meters. In the late 1980s and early 1990s, applications of GPS were largely for finding a user's current position, and were largely restricted to special uses such as marine navigation, hiking navigation, and a few similar special purposes. However, in the early 1990s, devices began to appear that had the capability to store a few hundred track points, where track points are defined as the discrete locations measured by the GPS device.

As the result of a conference held in Irvine, California, in 1996 (Transportation Research Board, 1996), the US Federal Highway Administration commissioned a proof-of-concept study on using GPS to track people's movements as a supplement to a household travel survey (Wagner, 1997). At this time, a GPS device that could store large quantities of data did not exist. Therefore, the test involved creating a device by attaching a GPS antenna to a hand-held personal digital assistant (PDA), and programming the latter to store the position

information from the GPS antenna. In addition, the PDA was programmed to display a series of questions at the outset of each trip, to which the respondent replied by using a stylus on the touch screen. These questions were designed to obtain additional information about each trip that could not be measured by a GPS device. The devices so developed were designed to be placed in a car, plugged into the cigarette lighter/accessory socket, from which the power was obtained, and the GPS antenna was placed on the roof of the car, with wires leading around the passenger front door of the car. The PDA had to be turned on at the beginning of each trip, at which time the respondent would enter the requested data about who was driving, who passengers in the car would be, and what the purposes were of each person in the car. If, during the execution of the trip, a change was made, such as by making an intermediate stop, the driver of the car was requested to enter correction information into the PDA for the trip now just accomplished, and also to add in revised information for the next trip.

The principal problem with this methodology is that it is fairly burdensome on the driver of the vehicle, and is also subject to missing trips when the driver forgets or does not have time to enter the information and turn on the PDA at the beginning of the trip. Notwithstanding these problems, the experiment was considered a success, and provided the impetus for others to begin experimenting with using GPS in various ways to measure people's travel. Lee-Gosselin and others began using GPS with measurement of various attributes of vehicle operation (Ueno et al., 1999). In the Netherlands, work was done in developing a version of a GPS system with a PDA that could be carried around, and especially that might be used by a bicyclist (Drajier et al., 2000). The battery was rather heavy in the device developed in 1999, so it was not as effective as had been hoped. New battery technology may solve this shortly. At Georgia Tech, work was also underway to develop new versions of a GPS/PDA device (Guensler and Wolf, 1999) that could also be carried outside a vehicle. Finally, at Louisiana State University, efforts were made in a different direction. This was to get rid of the PDA and minimize the actions required from the respondent, by using a passive GPS device with a prompted-recall survey following downloading of the stored GPS data (Bachu et al., 2001). This work has subsequently been extended with the acquisition of new GPS devices that are custom made for transport purposes (Stopher, 2001; Stopher et al., 2001, 2002a).

## 2.1. *Types of GPS device*

We classify GPS devices for use in these applications into two basic types, with the possibility of considering a subtype of one. The basic two types are active GPS and passive GPS. Active GPS devices are those that require the user to turn the unit on and enter data at the start of the trip or at some other point in the trip. An active

GPS unit consists of a GPS antenna/receiver connected to a PDA or some other type of data device that allows the user to enter data through a keyboard or stylus, and is able to display questions for answer by the respondent. Active GPS units may require the user to turn them off, or may turn off automatically either when the ignition is turned off, or when movement ceases for a long enough period of time. The original Lexington experiment used the first of these types of devices, and others continue in use. An example of an active GPS unit is shown in Figure 1.

The second type of GPS device is a passive device, in which the user does nothing. The passive GPS device consists of a GPS antenna/receiver connected to

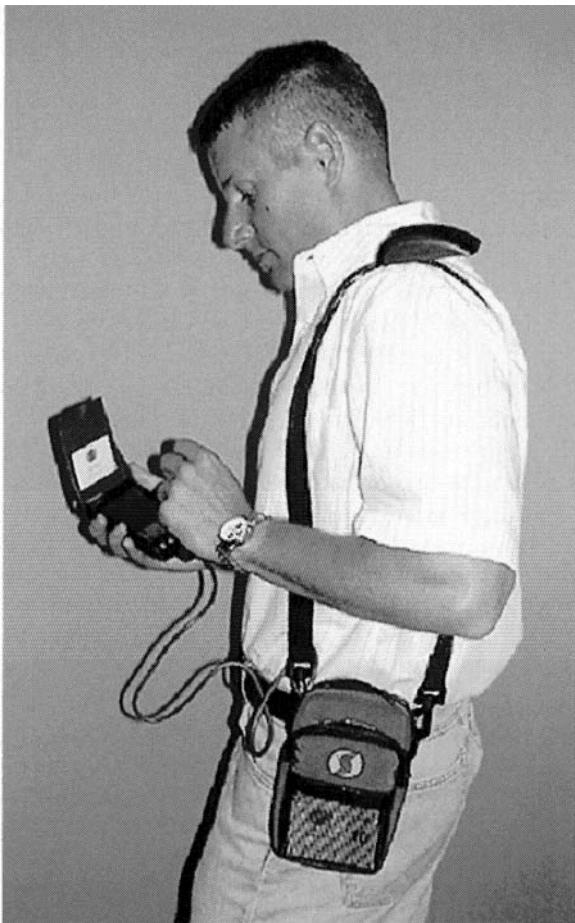


Figure 1. Example of an active GPS device.

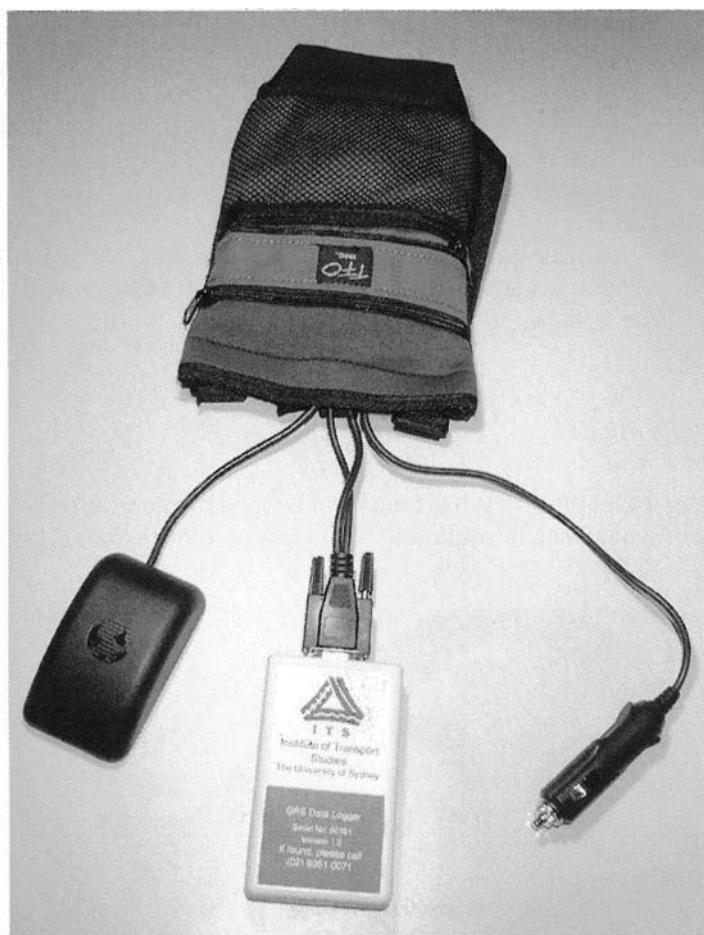


Figure 2. Example of a passive in-vehicle GPS device.

some type of passive data storage device. Data received by the antenna are automatically stored in the device. The device is turned on automatically at the beginning of travel, or is permanently on. If the device turns on automatically, then it also turns off automatically. With this type of device, the user is not required to do anything except make sure the unit remains in position and has power. An example of this type of device is shown in Figure 2.

There is a subtype of device that may be considered to be a partially passive device. This is a device that requires the user to turn it on, but no other action is required. This type of device was used by Bachu et al. (2001).

Most of the successful GPS devices in current use are still tied to a private vehicle, because of the need for a power source that can provide sufficient power to run the GPS antenna. Therefore, use of the cigarette lighter socket in a car remains the principal source of power for these devices. However, there are new developments in battery technology that appear likely to provide a wearable device that is very lightweight. A prototype developed for the Institute of Transport Studies at the University of Sydney has a battery that will run for up to 34 h and weighs as little as 170 g. In fact, the entire assembly of the battery, antenna, recording and logging device, a bag, and all cabling weighs a total of 710 g. This device is shown in Figure 3.

## 2.2. *What GPS can do*

GPS provides a capability that has long been needed by the transport profession. Depending on what is set, information can be recorded from once every second to



Figure 3. Example of a prototype passive wearable GPS device.

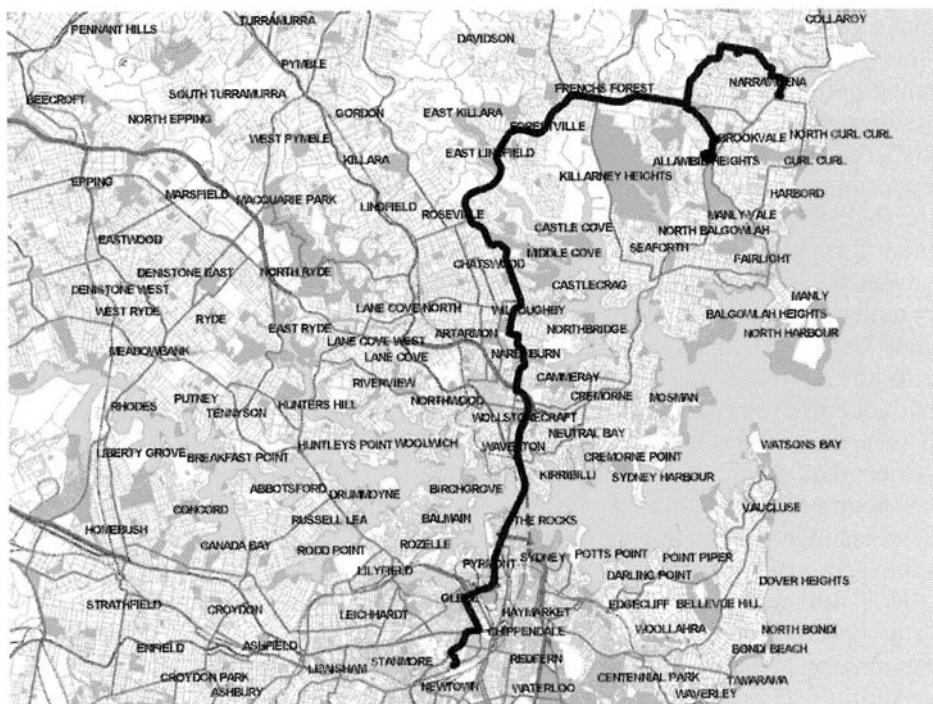


Figure 4. Track point data from a GPS device.

multiple seconds of the position and time of the GPS antenna. Position is recorded as the latitude and longitude of the position where the GPS antenna receives the signal, and is recorded to the nearest 1/100 000 of a degree. At the same time, the time at which each record is obtained is recorded in coordinated universal time (UTC)<sup>a</sup>, which used to be known as Greenwich mean time. This is recorded to the nearest second. In addition, information may be provided on the UTC date, the speed of movement (determined from Doppler computations), the heading, the number of satellites in view, and the horizontal dispersion of precision, which

<sup>a</sup> “UTC defines the universal time zone, to which all the other time zones are relative. It observes no daylight saving time (summer time). UTC is sometimes adjusted with “leap seconds” so that the difference between UTC time and the Earth’s rotational time does not exceed 0.9 seconds. The offset between UTC and TAI is always an integral number of seconds. The last leap second added to UTC was on December 31st 1998 at 23:59:60 UTC (the last minute of 1998 had one extra second). ... There are three important time standards: UTC Coordinated Universal Time, civil time, the one most often used by “ordinary” people; UT Universal Time, based on the Earth’s rotation, often used in Astronomy; TAI International Atomic Time, based on atomic clocks.” (Thorsen, 2002.)

measures how the satellites are clustered in the sky at the time of the observation. This provides unprecedented accuracy on where a person or a vehicle is at any point in time. Current commercially available GPS antennas and recording systems are capable of providing position accuracy to within about  $\pm 10$  m. Thus, GPS can provide very accurate data not only on the location of the origin and destination of each trip made, but also on the time of departure and arrival, and the route taken. It requires an accurate GIS map of the region, on which to overlay the information, in order to be able to determine more detail about the origin, destination, and route taken. An example of the track point data that can be obtained from a GPS device without differential correction, and presented over a good GIS base map, is shown in Figure 4.

With GPS that is tied to a vehicle for power requirements, the above data can be obtained from any car trip made. In the application of this procedure some studies have provided one GPS device per household, asking that it be placed in the vehicle that is used the most by the household. In other studies, one GPS device has been provided for every vehicle in the household. As wearable devices are developed, it will become possible to track all travel made, whether in a private vehicle, in a public transport vehicle, on foot, or by bicycle.

However, time and position is basically all that the GPS device can record. It cannot record mode, purpose, number of accompanying persons, out-of-pocket expenses, etc., – all of these must be determined through other means.

Another important aspect of GPS devices, however, is the completeness of the information obtained. In the past, transport analysts have had to depend on household travel surveys as the means to determine how much travel is undertaken by members of a household. While there have been many improvements over time in the techniques used to measure travel, it remains a fact that people forget to report some of their travel. As a result, it has always been believed that household travel surveys underestimate the number of trips made. With GPS devices, the only way in which the in-vehicle device will fail to collect travel is if it is not installed or is temporarily unplugged by the user. A wearable device will fail only if it is not carried with the respondent. Thus, there is not only the potential for GPS to collect data that are much more accurate as to location and time of travel, but also there is the potential to collect more complete data on the number of trips made. This is discussed further later in the chapter.

### *2.3. What GPS cannot do*

As noted above, GPS cannot collect such data as the trip purpose, the number of accompanying persons and their purposes, out-of-pocket costs, or mode used. Therefore, to obtain this information, there must be some form of supplementary data collection. As noted, this can take one of two forms. The GPS device can be

attached to a hand-held computer or PDA, and the respondent is asked to enter information during the travel episodes on the additional attributes that are desired. Alternatively, the GPS data can be downloaded, and maps created that are then used as the prompts in a prompted-recall survey.

In experiments with the prompted-recall survey, Stopher et al. (2001, 2002a) have found that people are able to recall the desired details of their travel up to at least 2–3 weeks after it was recorded, once they are shown maps of their travel or are given a tabular presentation of their travel. Beyond 3 weeks, some trips cannot be recalled in sufficient detail. In asking for details of trips made, the prompted recall also permits requesting if there are trips that were made that did not get recorded. Experience suggests that most of these trips are lost either through power failure to the unit or because of problems of signal loss, which we discuss further in the next section. The other source of loss is an undetected trip end in the middle of what has been identified as a trip in the algorithm for processing the track point data.

### 3. Processing GPS data

One of the principal challenges that arises for using GPS data to augment household travel survey data is to process the data collected into discrete trips that can be analyzed. There is little difficulty in determining that a trip has ended when a long period of time elapses between the end of one trip and the start of the next trip. Periods of time from tens of minutes to several hours are obvious as the ending of one trip and the beginning of another one. More difficult is the detection of a trip end that is very short, such as may occur when filling the car with petrol, dropping off or picking up a passenger, and other short duration activities. The problem becomes particularly difficult when the length of the stop may be equal to or less than that associated with stops in traffic, and when the ignition is not turned off. It is increasingly common to find traffic signals with 120 s cycles and longer, for which a car may wait in a queue for as long as 90 s without moving. It is also not uncommon to find that an unsignalized turn across a busy road can involve waiting times as long as 3 min or even 4 min.

Experience suggests that the use of an elapsed time of 120 s without significant movement is a reasonable rule in most urban areas as the threshold to define a trip end. The amount of movement permitted in this time must be small, but the position recorded by a GPS device is never static, because of the way in which GPS works. Stopher et al. (2002b) used a rule of 120 s without significant movement (defined as more than 10 m in any direction for each track point), and found this to identify 95% of actual trip ends and to identify spurious trip ends only about 3% of the time. Spurious trip ends are those that are suggested by the algorithm, but which are actually not trip ends, but long waits in the middle of a trip. Trip ends are

missed, when the trip end takes less than 2 min, which appears to occur less than 5% of the time. These spurious and missed trip ends can be partially corrected by visual examination of the data, and partially by a prompted-recall survey.

Another situation that needs to be looked for in GPS data is where a trip might end with a short stop, but is followed by the vehicle returning along the reverse direction to that in which it was moving prior to the stop. This is the situation that would arise if a driver were to drive into a street to pick someone or something up, or drop someone or something off, and then return in the opposite direction. What is required here is to check to see if, within a prescribed amount of time, the trip appears to continue in the opposite direction. In situations where a trip end has not previously been identified, but that within 120 s or less, the direction of movement is opposite to that which it was before, a potential trip end should be expected at the furthest point reached in the initial direction.

### *3.1. Problems with GPS data*

While the above rules will detect many trip ends, there are two major problems that arise in GPS data. The first is signal acquisition and the second is urban canyons, tunnels, and other similar situations. Most GPS antennae in use today have a rated signal acquisition time of about 15–45 s. However, this assumes that the GPS device is stationary for at least this amount of time after the power is turned on. The in-vehicle GPS devices that are the subject of this chapter are usually powered from the car ignition, and will not be on until the engine is switched on. Most people do not wait for 30 s before driving off after starting the engine. This is one advantage of the PDA-based systems, because, when the driver is asked to respond to several questions before leaving, the GPS antenna will usually have warmed up and acquired position before the driver leaves. At the same time, this is clearly a disadvantage of these devices, because there will be many occasions when the driver is unwilling to delay starting for 30 s.

GPS antennae have a warm start and a cold start. A warm start occurs if the device is reactivated within a short period of time (ranges from an hour to 2–3 hours), in which case the device is programmed to assume that the last position recorded is the location where the device is now. It then acquires satellite signals and begins to record its new position within usually about 15–45 s. A cold start occurs when the device has not been activated for a longer period of time. In this case, the last position recorded has been purged, and the device must reinitialize its position. This means that the device will take significantly longer to acquire position, because it not only needs to acquire the satellite signals but now needs to find a new starting location. This cold start may be prolonged.

When the GPS device is put in motion almost immediately that it is powered, it takes significantly longer to acquire position, unless it is in a warm start mode. If

the vehicle is driven off almost immediately the ignition is turned on, then it may take anywhere from 15 s to 4–5 min to acquire the signal, depending on the speed at which the vehicle is driven, whether any stops are made in that first short period of driving, and whether there are tree canopies or tall buildings close by. Stops allow for quicker signal acquisition, while tree canopies and tall buildings can cause further delays. Unfortunately, in 5 min, a car can travel 4–5 km, which may be the entire trip. In this case, it is possible for the entirety of short trips to be missed altogether by the GPS device. In other cases, there may be a gap between the last known location of the vehicle and the first point that is picked up when the new trip is undertaken. Stopher et al. (2002b) have developed software that corrects for missing data at the beginning of a trip, by assuming that the last point at which the vehicle ended a trip is the point at which the new trip began. Data are then interpolated from the underlying GIS map to add distance and time to the trip and to impute the likely start time of the new trip. However, this cannot correct for an entirely missed trip.

The second problem is where the driven route takes the vehicle between tall buildings, or into a tunnel, or through an area with a heavy tree canopy. Under these circumstances, the signal is often lost, because the GPS device can no longer “see” sufficient satellites to be able to fix position. In addition, the satellites in view are likely to be arranged in line, in which case the precision of the position fix is compromised. The result is often that either there are points missing over some considerable distance, or that points are widely scattered around the probable path. Again, Stopher et al. (2004) have developed an algorithm to handle these situations, impute the likely path during lost or compromised signals, and determine if there was a likely stop in the region where the signal was lost or compromised. This is based on examining the speeds immediately prior to signal loss and immediately following reacquisition of the signal, and also on estimating the length of the probable path through the affected area. A related problem in the same circumstances is the reception of bounced signals from the GPS satellites. These bounced signals off tall buildings can result in positions being shown, through an urban canyon, that may be far away from the true position. An example of this is shown in Figure 5.

### 3.2. Accuracy of GPS

For many purposes, even with the removal of SA, there can still be a need for the differential correction of the GPS positions, when a high level of positional accuracy is required. However, in the case of using GPS to determine origins and destinations of trips, and the route taken, differential correction is hardly necessary. Most modern GPS devices are capable of recording data to within  $\pm 10$  m or less. Assuming that the underlying base map is properly projected and



Figure 5. Scatter of points in an urban canyon.

accurately laid out, this is sufficient for the points to usually lie on the correct side of the road centerline. In fact, all that is generally required in these applications is that the GPS track points distinguish between one road and another, for the purposes of identifying the route, and that the end of the trip is recorded accurately enough to pinpoint a small range of addresses where the trip end was likely to have occurred. This can be achieved without any differential correction.

Estimation of travel times, distances, speeds, and acceleration are also sufficiently accurate from the uncorrected GPS records for virtually all planning applications. The information provided by uncorrected GPS is far beyond the accuracy of anything that has previously been available to transport planners.

### 3.3. Wearable GPS devices

The major thrust of the work to date on GPS relating to measuring personal travel has been with in-vehicle devices. These devices have been relatively easy to develop, because of the availability of a power supply, and the ease of having potential survey participants place the unit in their car and leave it there while measurement takes place. However, in-vehicle devices are obviously not capable of collecting data on walking trips, bicycle travel, and travel on public transport

vehicles and taxis. What is needed to measure other traveling events is a wearable GPS device. There have been two limiting factors: weight and bulk of the device, and power source. Some early wearable devices were developed around 1999 (Drajier et al., 1999), but these were quite heavy and bulky, and battery life was not very long. More recently, some wearable devices were developed in 2001 in the USA that ran on three D-type cells. The life of these units, if left switched on at all times, was about 64 h. At the end of that time, the batteries were sufficiently discharged that no further recording would succeed. The batteries made the unit fairly heavy and quite bulky.

Wearable devices are now under development that use rechargeable batteries. One version has a rechargeable battery that has a life of about 48–72 h, but is still about as bulky and heavy as the device using D cells. It also has a recharger that is fairly bulky; and it is not practicable for respondents to recharge the unit during use, notably because it requires the battery pack to be removed and slid on to the recharger base. However, a prototype unit that is expected to be available very shortly has a much smaller rechargeable battery with a life of about 32 h, and requires nothing more than a very simple plug-in charger, which can be done with the unit still in its carrying bag. This device weighs less than 1 kg and is no more bulky than two mobile phones. All of these devices have in common a small carrying bag that contains the logging unit and the battery pack, and a cable threaded through the shoulder strap of the bag, leading to the GPS antenna, which is mounted on the top of the shoulder strap, as shown in Figure 6.

Apart from the technology of a wearable GPS device, there are some data issues that are just beginning to be explored and analyzed. Even though these devices resemble the in-vehicle ones, some new data issues need to be dealt with. Included among these are more serious problems with urban canyons, resulting from the partial blocking of the view of the satellites from the antenna by the human head. For example, a person walking along a street with reasonably tall buildings close to the footpath, and the antenna on the shoulder closest to the buildings, creates effectively an urban canyon effect, where the satellites in view are restricted to those overhead and immediately in front and behind the pedestrian. An interesting challenge is to develop algorithms that can determine which mode is being used at each point along a trip. Work is progressing on this at several research locations around the world, but no results have been published as yet. It purports to present some interesting challenges.

#### 4. The future of GPS

There are exciting developments that appear likely in the near future for this technology in application to measuring personal travel. First, there is the continuing miniaturization of GPS antennae, which should result in the potential

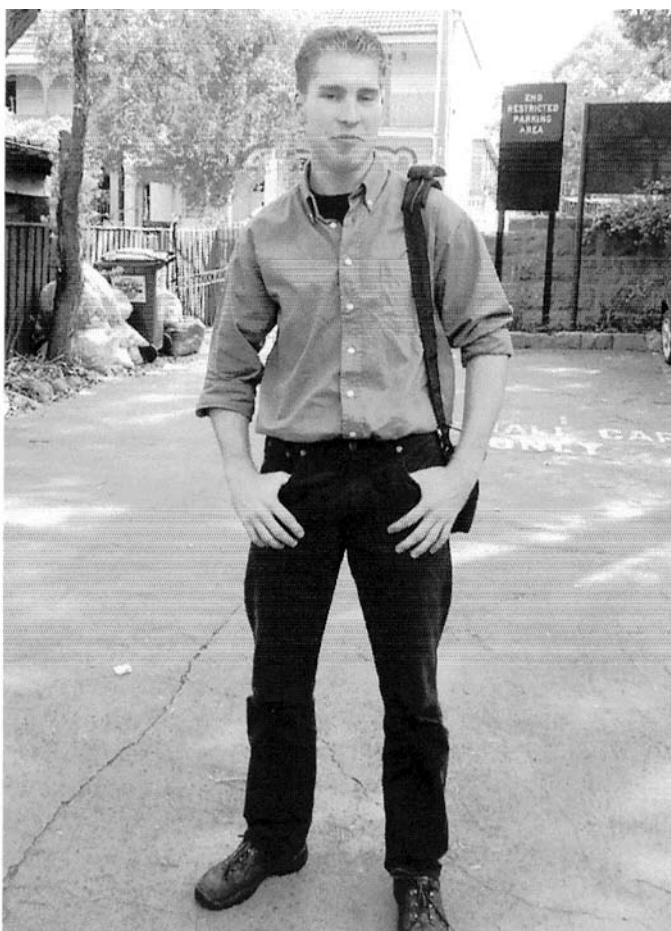


Figure 6. Wearing the small, lightweight prototype GPS device. Note that the GPS antenna is positioned on the top of the shoulder, affixed to the strap.

to create wearable devices that will be much more acceptable because of size and weight. This miniaturization may also make it possible to place an antenna in something that will allow better positioning for reception of signals. Second, there is the possibility of eventually being able to link GPS to a mobile telephone technology, which would permit the position to be obtained from the cells of the mobile telephone system, through triangulation. Such positioning with a mobile telephone is already possible, although its accuracy depends largely on the size of cells used in the system. In many areas, these cells are rather large so that position

information is insufficiently precise. However, future developments may make it possible to add positioning from a mobile telephone system, which would be able to take over when the satellite view is obscured, such as in urban canyons, inside shopping centers and other buildings, and in tunnels and vehicles that prevent the position from being obtained.

Third, there are prospects for further reductions in the size and weight of the battery packs for wearable devices. As GPS antennae are made smaller, their power requirements may also be reduced. Together with improvements in the power characteristics of rechargeable batteries, development of smaller and lighter batteries with longer life, the potentials for wearable units should be substantially improved over the present.

#### *4.1. Privacy*

Surprisingly, experience to date with in-vehicle GPS devices, and even some limited forays into the use of wearable GPS devices, has not raised significant issues of privacy. Such issues were anticipated by researchers and others working in this area, but the public seems, to date, to be more intrigued with the technology than concerned about privacy issues. Of course, individuals are probably quick to realize that, if they embark on travel that they do not want to be known, in-vehicle users simply need to unplug the GPS device for the duration of the travel, while users of wearable devices simply do not wear the device at such a time.

In many of the applications to date, fairly significant incentives have been offered (e.g. US \$50 per household in Lexington in the original proof-of-concept test, and incentives of US \$25 per household in some other recent surveys), so that the size of the incentive may have been sufficient to counteract any concerns over privacy. Nevertheless, in recent trials in Australia no incentives were offered, and issues of privacy have still not emerged in any obvious or significant way. Further, large-scale tests are needed to determine if this is a general result, or simply an issue of novelty at this point.

#### *4.2. Respondent burden*

One of the principal claims made by proponents of passive GPS devices is that these devices reduce substantially the level of respondent burden of this type of survey, particularly compared with the active GPS device. The latter requires data entry by the respondent for each trip undertaken, and for each destination visited. However, the development of the wearable GPS device poses some potential issues relating to respondent burden. At this time, it is too early to tell how significant this issue is and to what extent it can be reduced in such surveys. In

contrast to the in-vehicle devices, the wearable device requires the respondent to pick up the device and put it on each time he or she prepares to leave a location. Remembering to do this may be considered burdensome. In current versions, the respondent must also generally remember to plug the device into its recharger at the end of each day, adding a further burden. In addition, there is a literal burden, in terms of the need to carry the device around all the time while traveling. Future developments promise to improve this last aspect through potential reduction in size and weight of the antenna and the battery. Also, the recharging issue may be reduced, if future batteries, now in developmental stages, have longer life and require less frequent recharging.

## **5. Conclusions**

Experiments to date with GPS devices have shown that these are viable as a means to obtain detailed data on the trip movements and patterns of individuals in the general public. This is proving to be an increasingly promising procedure to determine both more accurate data on origins, destinations, and trip lengths, and also to provide route choice data that have not heretofore been available. There still remain questions to be answered relating to the most effective ways to deploy such devices, privacy, and respondent burden. These issues are likely to become more significant as the technology allows a shift from in-vehicle to wearable devices.

In the author's opinion, the passive GPS device is clearly superior, particularly when coupled with a prompted recall survey. The proven ability to develop software that provides efficient analysis of the vast amount of GPS data that can be obtained through these devices, and produce user-friendly reports on the travel undertaken, is clearly a major asset to the further implementation of passive GPS units. Furthermore, the development of strategies to repair faulty data from the GPS device, such as that caused by warm-up periods and urban canyon effects, makes the technology increasingly valuable. It is possible that future GPS developments may also reduce the severity of these problems.

It is unlikely that GPS will replace the standard household travel survey within the foreseeable future. It seems likely that the time requirements to deploy and retrieve the devices, analyze the data, and conduct supplemental surveys, coupled with the expense of the units themselves, are likely to make it prohibitive to conduct, within a few months, a sample survey of 3000 or more households using such technology. At this point, GPS appears to be most worthwhile when applied to a subsample of households within a standard household travel survey. Used in this way, GPS devices provide a wealth of data on route choices, and also provide a means to assess the accuracy of the companion household travel survey. Methods to take the GPS results and develop factors to apply to the standard household

travel survey results are under development, and would permit the subsample to be used to adjust the full sample to take account of those trips that people tend to omit from any form of household travel survey reporting. In addition, the GPS results may be able to be used to determine typical drive cycles for estimating emissions impacts of car travel, and also to track patterns of use of public transport at a level of detail never before considered possible.

## References

- Bachu, P.K., T. Dudala and S.M. Kothuri (2001) "Prompted recall in a GPS survey: a proof-of-concept study," in: *Transportation Research Board Annual Meeting*. Washington, DC.
- Draijer, G., N. Kalfs and J. Perdok, (2000) "GPS as a data collection method for travel research," in: *Transportation Research Board Annual Meeting*. Washington, DC.
- Guensler, R. and J. Wolf (1999) "Development of a handheld electronic travel diary for monitoring individual trip making behavior," in: *Transportation Research Board Annual Meeting*. Washington, DC.
- Stopher, P.R. (2001) "Using passive GPS as a means to improve spatial travel data," in: *Proceedings of the 8th World Congress on Intelligent Transportation Systems*. Sydney.
- Stopher, P.R., P. Bullock and S. Greaves (2001) "Using passive GPS as a means to improve spatial travel data: further findings," in: *Proceedings of the 23rd Conference of Australian Institutes of Transport Research*. Melbourne.
- Stopher, P.R., P. Bullock and F.N. Horst (2002a) "Exploring the use of passive GPS devices to measure travel," in: K.C.P. Wang, S. Medanat, S. Nambisan and G. Spring, eds, *Proceedings of the 7th International Conference on Applications of Advanced Technologies to Transportation*. Reston: ASCE.
- Stopher, P.R., P. Bullock and Q. Jiang (2002b) "GPS, GIS and personal travel surveys, an exercise in visualisation," in: *25th Australasian Transport Research Forum*. Canberra.
- Stopher, P.R., Q. Jiang and P. Bullock (2004) "Using GPS to measure travel times, congestion, speeds and performance," *Computer Aided Civil and Infrastructure Engineering* (in press).
- Thorsen, S. (2002) *Time zone abbreviations*. Stavanger: Thorsen tid og kalender (<http://www.timeanddate.com/time/abbreviations.html>).
- Transportation Research Board (1996) "Household travel surveys: new concepts and research needs, Irvine, CA," in: *Transportation Research Board Conference Proceedings No. 10*. Washington, DC: National Academy of Sciences.
- Ueno, M., N. Noël, S.T. Doherty, M.E.H Lee-Gosselin, F. Théberge and C. Sirois (1999) "Extending the scope of travel surveys using differential GPS," in: *ION-GPS 1999*. Nashville.
- Wagner, D.P. (1997) *Global positioning systems for personal travel surveys: Lexington area travel data collection test*, Report. Washington, DC: Battelle Transportation Division, US Department of Transportation.

*Chapter 25*

## GPS AND VEHICULAR TRAVEL

GEOFF ROSE

*Monash University, Clayton*

### 1. Introduction

The development of the satellite-based Global Positioning System (GPS) has had profound implications for many fields, of which transport is but one. While originally developed by the US military, the availability of the signals broadcast from the constellation of GPS satellites circumnavigating the earth has spawned a variety of non-military applications, many of which are marketed directly to consumers. While less well known, the Russian Global Navigation Satellite System (GLONASS) provides similar capabilities, but limited satellite availability and reliability problems mean that it is difficult to use as a standalone location system (Khoo and Luk, 2002). The European Commission and the European Space Agency are now developing Galileo, a satellite-based system that, unlike the US and Russian competitors, will be under civilian, not military, ownership (Khoo and Luk, 2002). The availability of satellite-based location systems seems assured under these developments, and this provides confidence in the long-term future of applications that rely on this technology.

Determination of a vehicle's location is a fundamental prerequisite for a host of intelligent transport system (ITS) applications, including delivery of in-vehicle information and emergency services. Importantly, these applications are not limited to the private car but are also relevant to freight and public transport operations. While a number of location technologies can be employed in these applications, GPS has the advantages that it is ubiquitous, quite precise, and the size and cost of the receiver units are reducing. The last point is important because it facilitates portable applications that rely on battery power. However GPS does have its disadvantages. For example, signal loss can occur in urban canyons, under heavy foliage, in multi-level car parks and underground. These signal loss problems need to be managed in operational systems.

This chapter reviews the range of applications where GPS is currently having, or will have, an impact on vehicular travel. We start by introducing a framework that identifies the key technology links that are needed to build applications around

the availability of the location information provided by GPS. The compatibility between those applications and various vehicular travel sectors, specifically private motorists, freight, and public transport, are then identified. Individual applications are then considered in more detail. The final section explores a number of critical and unresolved issues that will be influential in determining the future impact of GPS-based systems on vehicular travel.

## 2. Key technology links and applications

Table 1 highlights how key technology subsystems are packaged into applications for particular travel sectors. While the applications considered in this chapter rely on the ability to locate the vehicle, they also draw on a range of other technologies. The capacity to locate the vehicle is an essential component of the systems considered here, but it is important to recognize that GPS is but one approach for providing location information. Other options include radio signal trilateration, which uses multiple receivers to track and locate a single transmitter, mounted on the vehicle, which is emitting a time referenced signal at known frequency. GPS operates in reverse to trilateration, with a single receiver located in the vehicle while three or more transmitters (the satellites) are emitting signals from known locations. "Proximity signposts" generally consist of some sort of identification tag that is used to determine the proximity of the vehicle to a particular location (the "signpost"). The tag may take the form of a radio transmitter, and the signpost then requires some way of reading or communicating with the tag. This approach has been used to develop a system for automatically collecting travel time information from a fleet of probe vehicles in Sydney (Longfoot, 1991). Another emerging area, which has the potential to provide a more ubiquitous capability, is the location of mobile or cellular phones through analysis of the signals transmitted between the handset and the base station (Ygnace et al., 2001). In another variant, GPS units are also being installed in mobile phones to enable the location of the handset to be determined automatically when an emergency call is placed. While acknowledging the availability of alternative location technologies, the emphasis in this chapter is on applications that rely on GPS, since these are seeing increasing application.

Mapping systems are central to in-vehicle navigation applications and many other applications, as noted in Table 1. These can provide road-segment-related information or contain richer information about land use or destinations. The latter can be useful for navigating to a point of interest (e.g. to find the nearest florist or a major shopping center). These mapping data are usually stored in the vehicle – usually on a CD-ROM. However, there are issues related to the cost and currency of that information. Some applications are now relying on mapping data stored off-board, at a service center, and relevant information is downloaded to the

**Table 1**  
Vehicular travel applications for GPS

Application	Key technology subsystems					Vehicular travel sector		
	Location (e.g. GPS)	Mapping	Communications	Vehicle monitoring	Payment	Private motorists	Freight/heavy vehicles	Public transport
Remote monitoring of vehicle location	✓	✓	✓	✓		✓	✓	✓
Arrival time information	✓	Possibly	✓			✗	✗	✓
In-vehicle navigation	✓	✓	✓ (for dynamic IVNS)			✓	✓	✓ ("dial-a-ride")
Intelligent speed adaptation (ISA)	✓	✓		✓		✓	✓	✗
Advanced driver assistance systems	✓	✓	✓	✓		✓	✓	✗
Electronic payment and charging	✓	Possibly	✓		✓	✓	✓	✗

vehicle on an as-needed basis. This makes possible “pay per use” charging systems, thereby avoiding the potentially high fixed cost associated with the purchase of vehicle-based hardware and maintaining the full map database in the vehicle.

Communications systems are also important, particularly for the provision of dynamic, or real-time, information into the vehicle. This is important for private vehicle applications as well as for freight and public transport operations. To support vehicular-based applications these need to be wireless systems, such as radio or mobile/cellular telephone systems (e.g. the Global System for Mobile Communications – GSM). A GSM platform makes possible two-way communications, to and from the vehicle. In this way, data can be downloaded, for example real time speeds in the road network, while the vehicle can also pass data back to the control center. In the latter case the vehicle itself may be acting as a probe, passing time-stamped location data back to the central control center from which the prevailing network speeds are determined.

Applications may rely on links to a range of vehicle-monitoring systems. Some of these are used to overcome the limitations of GPS, as mentioned in the introduction. When there is a loss of signal, dead reckoning can be used to establish the vehicle location, and this relies on establishing the direction in which the vehicle is traveling (from an on-board gyroscope) and the distance traveled since the last known location fix (obtained through the odometer). Intelligent speed adaptation (ISA) systems provide feedback to the driver when the speed of the vehicle exceeds the prevailing speed limit, and this requires a link to the vehicle odometer. Mayday systems, developed to summon assistance in the case of an emergency, may link into other vehicle-monitoring equipment such as sensors that detect if the airbags deploy. In these systems, airbag deployment generates a call on the GSM phone to the service center that includes the location of the vehicle (provided by GPS) and an alarm that the airbag has deployed. The service center may then call the vehicle to try and establish communication with the driver/occupants. The service center then advises emergency services accordingly, and can pass on precise location details for the scene of the emergency. These mayday systems are particularly valuable in remote regions where morbidity and mortality can be reduced by improving response times over reliance on manual reporting. In public transport applications, passenger loads may be monitored, while in freight applications a diverse range of sensors may provide information on the load (weight and distribution across axles, stability), the engine (fuel flow, maintenance thresholds), and safety equipment (brake conditions).

Finally, payment systems may be based in the vehicle, in the case of a stored-value “smartcard,” or processed off-board through a charging system that relies on data about the location and use of the vehicle. There is growing worldwide interest in electronic road pricing, and while to date only limited applications rely on GPS there are indications that GPS-based systems may see increasing application in the future.

The right-hand columns of Table 1 highlight the relevance of individual applications to particular travel sectors, specifically private motorists, freight, and public transport. Many of these applications were initially developed for the freight sector before they became available to private motorists. While perhaps a late starter, the public transport sector is now seeing increasing deployment of automatic vehicle location and arrival time information systems. Each of the applications identified in Table 1 is considered in turn in the following sections.

### 3. Remote monitoring of vehicle location

By combining the precise location information provided by GPS with a wireless communications capability such as GSM, a range of safety and security applications are possible. In a simple form these can track and locate stolen vehicles (Royal Automobile Club, 2003), or by linking to other vehicle-monitoring equipment, such as an indication that the air bag has deployed, emergency notification systems (known as mayday systems) are provided. One of the more successful of these is the OnStar system, operated by a subsidiary of the General Motors Corporation in the USA. OnStar has over 2 million users in the USA, and currently receives about 500 air bag deployment notifications per month (George Washington University Hospital, 2002). Car rental companies can also use the same technology platform to monitor vehicle use and determine whether vehicles are taken on unsealed roads or driven across state or national borders. A further link to the odometer, or by using the speed data available through GPS, also provides a capability for vehicle rental companies to determine whether the vehicle is being driven at excessively high speed (Lemos, 2001).

Heavy vehicle operators and road authorities/regulators are potential beneficiaries of the real-time location monitoring capability that is provided by GPS. Freight companies use GPS tracking to monitor valuable or hazardous cargoes. The data may relate to vehicle or cargo location, with the latter of particular relevance in port operations where the ability to monitor cargo status can enable transport operators to plan pick up and delivery operations to reduce delays (US Department of Transportation, 2002). There are also possibilities for information to be provided to regulators in exchange for opportunities to improve productivity. In this context, the National Road Transport Commission in Australia is exploring the scope for an ITS application, built around the location information provided by GPS, to provide a basis for ensuring that heavy vehicles comply with their permitted operating conditions, ensuring that they operate how, when, and where they should (Koniditsiotis, 2002). Potential applications cover large or over-dimensional vehicles as well as hazardous loads. A pilot project operating in the state of New South Wales focuses on mobile cranes. By allowing

their fleet to be monitored continuously, the operators are permitted to adopt operating procedures that improve vehicle productivity. The location monitoring is conducted by an independent third party that automatically notifies the state road authority of any breeches in operating conditions.

Remote vehicle monitoring is not limited to freight applications but is relevant in the context of public transport operations as well. For example, school buses can be monitored with GPS to provide parents and school transport administrators with real-time information on the location of vehicles (Oloufa, 2003). This information can be made available to assist parents to schedule drop-offs or pick-ups to/from the bus to minimize waiting times, particularly during severe weather, or to assist transport administrators to respond to accidents.

#### 4. Arrival time information

For public transport operations, the location information provided by GPS can be used as the input for algorithms to predict the arrival time of individual vehicles at downstream stops. This provides the basis for displaying arrival times on variable message signs at the stops. The provision of real-time arrival information not only reduces the perceived waiting time for passengers but gives them the opportunity to undertake other activities in the time they know they have available until the arrival of the public transport vehicle.

It is appropriate to note that in the public transport context, GPS is not the only technology used to provide automatic vehicle location (AVL) data. However, Hounsell and Wall (2002) note that GPS is becoming the preferred AVL technology because of the features it offers in terms of flexibility, lack of need for roadside infrastructure, and improved accuracy with the removal of intentional signal degradation (known as selective availability).

AVL is commonly implemented along with automatic priority systems for public transport vehicles that aim to reduce the delays they experience at traffic signals. Knowing the location of the public transport vehicle enables its expected arrival time to be estimated and compared to the scheduled arrival time in order to make a decision about whether signal priority is required.

These systems have been demonstrated to increase ridership by about 10% in the case of bus operations in Finland (Lehtonen and Kulmala, 2002), while in London the system that provides arrival information at bus stops has produced a 1.5% increase in cash revenue on routes where the arrival information is displayed (ERTICO, 2003). The ridership impact will depend on the headway and reliability of the service reflecting the benefits to travelers through provision of information on arrival times and delay reductions through priority treatment at intersections.

## 5. In-vehicle navigation

Navigation can be defined as the process of directing the movements of a user from one point to another. Successful navigation is therefore dependent on establishing where the user is, where the user wants to go, and how to get there. Clearly, establishing location is a major issue in navigation, and hence GPS has an important role to play. GPS is but one of the technologies required for an operational in-vehicle navigation system (IVNS). As noted in the introduction, there are times when the GPS signal is lost, and so other sensors are required to determine the location via dead reckoning. While public transport services usually operate to a fixed-route, a “dial-a-ride” operation could benefit from an in-vehicle navigation system that directs the driver to the next pick-up or drop-off point, with passenger requirements (specifically pick-up points) relayed dynamically to the vehicle. An INVS can also be used for freight operations, particularly delivery services where it may be integrated as part of an automated dispatch system that may ultimately feed into a vendor-managed inventory (VMI) system designed to reduce inventory and transportation costs (Rabah and Mahmassani, 2002).

In-vehicle systems can be classified as either static (autonomous) or dynamic. The dynamic systems receive traffic information in real time, while the simpler systems operate independently, or autonomously, on the basis of static information. We will consider the static systems first.

Most navigation systems use a combination of dead reckoning, map matching, and GPS to determine the position of the vehicle. Map matching is undertaken to match the path of the vehicle, as determined from dead reckoning, with the digitized road map stored on a CD-ROM. The process involves the use of algorithms to compute changes in direction and distance traveled. When new data are received from the GPS and dead-reckoning functions, output coordinates are linked to known points on the digital road map. A CD-ROM drive gives access to CD-ROMs containing digital maps of road networks perhaps for different cities or prepared in different languages. Information on points of interest and useful services such as hotels, restaurants, service stations, and hospitals may also be available on the CD-ROM. There is scope to provide a lot more destination-relevant information, including video clips of hotel facilities, local attractions, etc., by changing from CD to DVD as the basic storage medium for an IVNS.

The user interface consists of a device for input of data and control commands and a small screen/monitor that is used for presentation of input data (desired destination), maps, distance to destination, etc. Turn-by-turn instructions are displayed on the screen and transmitted audibly.

The destination is entered by either recalling addresses from an address book or entering the address of the destination by selecting characters, one at a time, from an alphabetic list on the screen. Once the destination has been entered, the system calculates the best route, which is displayed on the monitor along with a

background map. It is common for the systems to provide the user with options to influence the character of the route selected by the navigation computer. Options typically include shortest time, shortest distance, most or least use of freeways, and avoid tollways.

The guidance may be displayed on the screen via a map that includes a symbol indicating the current location of the vehicle along with a combination of audible and/or graphical turn-by-turn advice. Under acoustic/voice guidance, for every turn, a synthetic voice informs the driver how far away the turn is and what direction to turn; then another verbal instruction is given as the vehicle approaches the turn. Besides spoken route directions, some systems also offer visual route information, either as simple directional arrows (pictograms), or combined map-pictogram displays. The voice output is the main navigation tool, while the pictogram provides a quick visual check of the verbal instructions. Studies undertaken in the USA (Peters and Mammano, 1995) concluded that turn-by-turn guidance information (whether presented orally, as a textual list, or by a graphic display) enhances the performance, usability, and/or safety when compared with alternatives that provide route information. Should the driver miss a recommended turn, or decide to change route, the IVNS calculates a revised route automatically. This allows the driver to concentrate on the task of driving, with the system providing ongoing navigation assistance.

Compared with cheaper paper maps, cost is obviously a major disadvantage for navigation systems. To make the IVNS more appealing to potential consumers, additional functions will be needed, and the integration of real-time and specific traffic information presents opportunities in this regard. There are a number of operational systems that are providing real-time traffic information. These systems depend on receiving real-time information on traffic conditions. A range of technologies are used for communicating the real-time data, including systems of beacons as used in VICS in Japan (Okamoto et. al., 1998), FM broadcasting as used in VisionAute in Paris (VisionAute, 2003), or mobile phone (GSM) links as used in Smartnav in the UK (RAC, 2003). While these systems can assist the driver under recurrent, or normal, congestion, they provide potentially valuable information during incidents (accidents, breakdowns, localized heavy demand due to a special event such as a football match, etc.). These systems may also rely on the GPS device to have the vehicle serve as a probe and relay dynamic travel time information to the control center (Nuttall, 1999).

To overcome the cost of the processor and map data being stored in the vehicle, there is a move toward systems where those components are stored remotely. In these systems, the location of the vehicle is still monitored by GPS, and the destination is communicated to the control center electronically via the in-vehicle device or by voice to a call center operator. Turn-by-turn guidance, based on dynamic travel time information, is then downloaded to the vehicle. Operational versions include the Tegaron Scout, which is available in Germany (Tegaron,

2003), and Smartnav in the UK (Royal Automobile Club, 2003). This opens the possibility of a “pay-per-route” option, which may reduce the cost of accessing dynamic travel time information for drivers who only need this information infrequently. Another innovation in the platform on which an IVNS is offered is that rather than relying on the same amount of in-vehicle equipment, some systems operate on a personal digital assistant (PDA), e.g. the Tegaron Scout. These completely portable units still rely on GPS for vehicle location and download all necessary dynamic navigation information from the control center, usually via GSM.

Traffic data are not the only type of information that motorists may wish to receive real-time updates about. Once the communications capability is provided for the vehicle to access real-time data, it would be just as easy to transmit information on car space availability in different car parks, petrol/gas prices at different service stations, etc. In the latter case, the driver may ask the system to provide directions to the nearest, cheapest petrol/gas station.

## 6. Intelligent speed adaptation

A relatively straightforward, at least in concept, extension of IVNS technology is the development of an ISA system. The aim of ISA is to reduce speeding anywhere on the road network (Oei and Polak, 2002). An ISA system performs three functions (Sundberg, 2000):

- measurement of vehicle speed;
- decision of a suitable speed;
- execution of speed adaptation.

The vehicle speed measurement is taken care of by the odometer installed in the vehicle. The determination of a suitable speed can rely on input from a GPS receiver. The map database (which could form part of an IVNS) can contain characteristics of each link in the network, including the declared speed limit. Once the location of the vehicle has been determined via GPS, and matched to a location on the map, the relevant speed limit can be readily determined. One trial in Sweden used beacons to transmit speed limit information to the vehicle (Oei and Polak, 2002). The speed limit can be classified as:

- fixed – reflecting the general speed limit on each road section;
- variable – the speed limit depends on specific road situations, e.g. near pedestrian crossings, or schools;
- dynamic – the speed limit depends on prevailing, real-time, weather, traffic, and light conditions.

Wireless communications (e.g. GSM) could be used to transmit revised speed limits due, for example, to local weather or roadworks. The form of the speed adaptation can vary:

- a flashing red light and increasing beeps, to signal to the driver that the speed limit is being exceeded;
- display of the prevailing speed limit along with a audible warning;
- a haptic device, whereby the accelerator pedal requires more pressure to depress it;
- an automatic speed reduction back to the speed limit.

It is common for the vehicles to have an override that is engaged using the step-down function under heavy acceleration, so that rapid acceleration is still possible in an emergency.

ISA systems are generating increasing interest, with a number of active research programs underway. One of the largest trials to date was conducted in Sweden in four towns (Martin, 2002). In each town the configuration of the system varied, and in each case they were voluntary ISA systems. On the basis of the results from the Swedish trials, it has been estimated that the reduction in killed or seriously injured accidents as a result of ISA installed in vehicles in built-up areas would be about 20% with a mandatory system.

There are a number of unresolved issues relating to public acceptance of ISA. For example, Oei and Polak (2002) express a concern that drivers will switch the system off unless it is dynamic. There is evidence that those in most need of ISA are likely to be the ones least likely to favor its introduction. In this context, UK research suggests that ISA will be more resisted by younger male drivers, least experienced drivers, high annual distance drivers, and those who drive as part of their work (Stradling, 2001). There is also an issue of whether these systems would become mandatory on all vehicles (e.g. speculation that this could be required to be fitted to all vehicles in the UK by 2010) or whether they would only be fitted to recidivist speeding motorists as one condition of a return to driving. Another issue is the chance of irritation on the part of drivers without ISA toward drivers in front with ISA (Oei and Polak, 2002) and the risk of increased risk-taking behavior of those following non-ISA equipped drivers. Despite reservations by some commentators, there is political support for ISA in Sweden, and building on the success of the trials conducted there, it may be the first country to offer all drivers an advisory ISA service (Martin, 2002).

## 7. Advanced driver assistance systems

Beyond ISA, the merging of speed advice and additional location-specific information has further potential to improve safety. Existing systems seem set to evolve to advanced driver-assistance systems (ADASs), which would aim to reduce collisions by improving driver awareness of their surroundings. Early applications will provide features such as (Hook, 2002):

- headlights that will automatically turn with the road;
- an adaptive cruise control that alters the speed of the vehicle to take account of road geometry.

Future applications will aim to provide the driver with an extended view of the environment in front of the vehicle through night vision systems, braking and stability assistance, intersection collision avoidance, run-off road detection/warning systems, and curve preview capabilities (Hook, 2002).

## 8. Electronic payment and charging

The availability of location data provided by GPS, in conjunction with other technologies, makes possible a range of payment and charging systems. Present road pricing and tolling systems rely on dedicated short-range communications (DSRC) from overhead gantries to establish the location where charges are to be levied. The use of GPS would facilitate a “virtual” tolling or charging cordon. Since the vehicle location is determined by GPS, the time when and location where the vehicle crosses a charging cordon can be determined. On a regular basis (e.g. once per month) that information could be downloaded to a charging authority. This could take place via a communications link, e.g. over a GSM mobile/cellular phone. A recent electronic road pricing study in Hong Kong concluded that vehicle location technology such as GPS was an attractive alternative to DSRC systems (Hong Kong Transport Department, 2001).

Singapore is well known for its leading-edge application of electronic road pricing (Luk, 1999), and is already exploring other options for a new-generation road pricing system. The existing system, which is based on DSRC from gantries, has a number of disadvantages, including an inability to respond quickly to worsening traffic conditions at particular locations, partly because of the time lag necessary to build and install the gantry. According to Juan (2002) the Land Transport Authority in Singapore is exploring a next-generation system that

- does not require physical roadside gantries;
- provides coverage throughout the country (island);
- has the flexibility to price roads on a per-passenger or on a distance-traveled or time-taken basis or other criteria.

The recently introduced heavy charging system in Switzerland provides one concrete example where some of those features desired in the next-generation system for Singapore are already operational. LSVA is a distance-based charging scheme that applies to all heavy vehicles (maximum laden weight exceeding 3.5 t) on all public roads in Switzerland (Oehry, 2001). The main goal of the system is to

internalize the external costs of freight transport. The heavy vehicle fee is charged on the basis of three key criteria:

- the number of kilometers traveled;
- vehicle weight (trailers are assessed together with the prime mover);
- the emissions category of the heavy goods vehicle.

Older and more polluting vehicles pay a per tonne-kilometer rate that is 40% higher than for the least-polluting vehicles.

The system relies on an on-board unit that records relevant trip data automatically. That data is downloaded periodically to a chipcard that is sent to the Swiss Customs Authorities by post or via email for regular billing. Distances are recorded via a connection to the tachograph, which is mandatory on all heavy vehicles in Europe. That is complemented with GPS and a movement sensor to ensure that the tachograph signal is not intentionally interrupted or falsified (Oehry, 2001). A DSRC link from overhead gantries is used to switch on the recording system. The driver uses a simple interface on the on-board unit to declare that a trailer is attached. A trailer detection device serves as a reminder to the driver, and also as the basis for an exception report to be included in the data if the driver does not declare a detected trailer. To facilitate straightforward external checking, the on-board unit is fitted with a series of lights that indicate whether the unit is operational and whether a trailer is declared.

Oehry (2001) reports three important types of impact resulting from the introduction of the system:

- adaptation of fleet composition, by replacing high-emission vehicles with new low-emission vehicles and better matching of vehicles to the load being carried;
- improved operational planning, extending to cooperation or merging of firms, to maximize load factors;
- greater attention to minimizing the distance traveled in Switzerland.

Oehry also notes an anticipated long-term shift from road to rail for long-distance freight traffic. This shift has not been observed in the short term, partly because funding from the scheme is being directed to expand the rail network and the operators have increased rates in view of the higher costs of road transport.

Another important application for GPS relates to distance-based charging for private motor vehicles. This can cover fixed costs such as insurance and registration fees. To date, the initial experience relates to motor vehicle insurance where the rate of exposure is changed from the vehicle-year to the vehicle-kilometer or vehicle-minute (Victoria Transport Policy Institute, 2002). While these systems can be implemented with monthly reporting of odometer readings, the systems under operation and testing rely on GPS to keep track of location,

timing, and distance information. The detailed location information provided by GPS makes possible a different basis for vehicle insurance risk assessment, which would depend on when and where the vehicle is driven. The “where” component could consider not only the geographic areas but also the facility types commonly used, e.g. freeways versus arterial streets. One insurance company in the USA has operated a usage-based insurance rating system called Autograph (Progressive, 2002), for a number of years. The system relies on GPS and mobile/cellular telephone technology to report usage statistics automatically to the firm. Progressive’s press release at the time the system was patented (Progressive, 2002) indicates that some consumers have saved 25% on their premiums, and others who live close to work, or have other cars that are used infrequently, are realizing even larger savings. In the UK, Norwich Union is trialing a “pay as you drive” system, which is based on GPS location technology and GSM for the communications system (Norwich Union, 2002). A total of 5000 motorists are being recruited as volunteers to participate in the trial, which will last until the end of 2004. The trial will provide the basis for establishing individual premiums based on how often, when, and where the vehicles are actually used.

While these systems are of interest to the individual motorist as one way of potentially reducing motoring costs, they also represent an opportunity to make motorists more aware of what were previously fixed costs of motoring. In that way, they are also of interest from the perspective of travel demand management, where a move to clearer usage fees associated with motor vehicle use can be expected to encourage motorists to internalize those costs each time they reach for the keys. The Victoria Transport Policy Institute estimates that a move in the USA to the anticipated average insurance cost of 6 US cents per mile (under 4 US cents per kilometer) would reduce vehicular travel by 10% or more (Victoria Transport Policy Institute, 2002). The benefits of pay-as-you drive vehicle insurance include consumer savings, economic efficiency, increased fairness, and affordability.

## 9. Unresolved issues

There are a number of issues relating to GPS that are yet to be resolved, and these are likely to have important implications for the development of GPS-based applications for vehicular travel.

### 9.1. Map database related

There are already issues related to the accuracy and timeliness of the databases that underlie in-vehicle navigation systems. At present, a delay of at least 6 months

between compiling source material and its availability to drivers reflects not only the technical process required to convert the underlying data into the different formats required by system manufacturers but also the commercial decisions made by the firms involved (Hook, 2002). If maps are to support safety-critical functions within the vehicle, then the data must be released in a timely fashion, and its accuracy also becomes critical.

The current updating process is to obtain a new CD-ROM and install it in the navigation computer of the vehicle. Future updating is likely to be via wired or wireless technology. Under the wired approach, drivers could obtain a download of the latest editions of maps at service stations, or overnight at home. The other option is wireless downloading of data to the vehicle, using a mobile communication system. However, the data transmission capacity (technically known as the bandwidth) of existing systems such as GSM will not allow the very large files to be downloaded quickly enough to be acceptable in terms of time or cost. Developments such as the third-generation (3G) mobile phone systems will provide much greater bandwidth, and some manufacturers are reporting success with intermediate technologies such as General Packet Radio Service (GPRS) (Hook, 2002).

For the non-safety-critical and dynamic systems there is also the issue of whether the data should be stored on the vehicle or at a control center and downloaded to the vehicle on an as-needed basis (as described earlier). Particularly in the context of dynamic in-vehicle guidance, this also makes possible pay-per-view options, which may stimulate market interest.

## *9.2. Human factor considerations*

The development of many ITS applications raises issues about implications for the human operator of the vehicle. Concerns can include the reactions of drivers when a system fails after they have developed a reliance on warning systems (e.g. ISA), cognitive overload arising from attempting to take on board too much information, or the distraction of information from IVNS. There is also the potential for risk compensation, whereby if drivers are forced to adopt ISA they will compensate for this mandated risk reduction by increasing risk through other behavior, for example reducing following headways. There is considerable scope for human factor research in the context of many ITS applications.

## *9.3. Willingness to pay*

While some organizations may originally have believed that in-vehicle navigation presented lucrative market potential, the consumer response has been subdued at

present prices. There are also issues of how much motorists are willing to pay for dynamic information when radio traffic broadcasts already provide information free of charge (Polydoropoulou et al., 1997). The evolution of off-board systems has the potential to reduce access costs, and this could be the technological breakthrough that will impact significantly on the costs of information, and therefore on the uptake rates for dynamic in-vehicle systems.

#### *9.4. Managing privacy*

By its very nature, GPS can provide very rich time-dependent location information. The scope for invasion of privacy is considerable. At the very least, the matching of location data with other data sources, e.g. credit card transaction or land use information, could provide opportunities for new niche marketing. In many cases operational systems are run by private-sector organizations, and the opportunity to generate additional revenue would be appealing. There are already government privacy principles enshrined in legislation (e.g. in Australia), and concerns have already been raised about the privacy implications of some ITS applications, particularly electronic tolling (Ogden and Doupe, 1998). As noted earlier, vehicle rental companies can monitor vehicle operation using GPS, and this has resulted in law suits in the USA where renters were subsequently fined by the company on the basis of the data provided by the monitoring system (Lemos, 2001).

#### *9.5. Public and user acceptance*

Apart from the technical issues associated with particular applications, there are also concerns about the extent to which motorists will accept applications such as ISA and some electronic payment options. These systems may fail because of inadequate attention to implementation issues rather than from technical problems *per se*. In this context there will no doubt be much to be learned about the introduction of electronic road pricing in London, which began in 2003, and the ongoing research work with ISA.

There is scope for greater understanding of routing algorithm features that would increase user acceptance of navigation systems. Alternatively, research could be directed at new generations of dynamic systems that could "learn" about the factors influencing a driver's route choice from their driving patterns, and use that information to provide a basis for route recommendations when guidance advice is sought by the driver.

## 10. Conclusions

The ubiquitous location-finding ability provided by GPS has already had major impacts on vehicular travel. Operational vehicle monitoring, mayday, and navigation systems are demonstrating that drivers value the features offered by these systems. Safety- and charging-related applications are likely to be the next frontiers for GPS in vehicles. The move to central data storage and downloading of only the currently needed information to the vehicle could well be the technological shift that brings down costs and stimulates greater market interest for in-vehicle dynamic navigation systems.

It is unlikely that technology will serve as a constraint on the deployment of applications that rely on GPS. Further research is needed on human factor implications, privacy management, and also public acceptance of systems such as ISA. Inadequate attention to those issues may be the greatest long-term threat to the growth of further applications in this area.

## References

- ERTICO (2003) *Countdown: real-time bus stop information in London, UK*. Brussels: ERTICO ([http://www.ertico.com/its\\_basi/succstor/countcon.htm](http://www.ertico.com/its_basi/succstor/countcon.htm)).
- George Washington University Hospital (2002) *General Motors announces advanced automatic crash notification*, Internet press release. Washington, DC: George Washington University Hospital (<http://www.gwhospital.com/p5779.html>).
- Hong Kong Transport Department (2001) *Feasibility study on ERP*, Final report. Hong Kong: Hong Kong Transport Department (<http://www.info.gov.hk/td>).
- Hook, P. (2002) "A more certain future," *Traffic Technology International*, Dec./Jan.:80–82.
- Hounsell, N. and G. Wall (2002) "Examples of new ITS applications in Europe to improve bus services," in: *Proceedings of the 81st Transportation Research Board Conference*. Washington DC.
- Juan, H.E. (2002) "A sustainable transport policy framework," in: *International Conference on Seamless and Sustainable Transport*, Keynote address. Singapore.
- Khoo, V.H.S. and J.Y.K. Luk (2002) "GPS technology and enhancement positioning developments: applications to road transport in Singapore," *Road and Transport Research*, 11:34–45.
- Koniditsiotis, C. (2002) "Intelligent Access Project (IAP) – feasibility," in: *Proceedings of the ITE ANZ International Conference*. Melbourne.
- Lehtonen, M. and R. Kulmala (2002) "Benefits of pilot implementation of public transport signal priorities and real-time passenger information," *Transportation Research Record*, 1799:18–25.
- Lemos, R. (2001) *Rental-car firm exceeding the privacy limit?* San Francisco: CNET News.com (<http://news.com.com/2100-1040-268747.html?legacy=cnet>).
- Luk, J.Y.K. (1999) "Electronic road pricing in Singapore." *Road and Transport Research*, 8:28–40.
- Martin, J. (2002) "After a 4 year trial – what the Swedes think of ISA," *Traffic Engineering and Control*, 43:376–379.
- Norwich Union (2002) *The pay as you drive insurance device*. Norwich: Norwich Union ([http://www.norwich-union.co.uk/products/insurance/motor/ap\\_as\\_you\\_drive/updates.htm](http://www.norwich-union.co.uk/products/insurance/motor/ap_as_you_drive/updates.htm)).
- Nuttall, I. (1999) "Visionaute sets out for France," *Traffic Technology International*, Feb./March:46–48.
- Oehry, B. (2001) "Inside LSVA – technical concepts of kilometre-charging in Switzerland," in: *Proceedings of the World Congress on Intelligent Transport Systems*. Sydney.
- Oei, H.L. and P.H. Polak (2002) "Intelligent speed adaptation (ISA) and road safety," *LATSS Research*, 26:45–51.

- Ogden, K.W. and P.J. Doupe (1998) "Privacy: a motorist's perspective," in: *Proceedings of the 5th World Congress on Intelligent Transport Systems*. Seoul.
- Okamoto, M., T. Yamamoto, K. Sakamoto and M. Tuge (1998) "Evolution of the VICS," in: *Proceedings of the 5th World Congress on Intelligent Transport Systems*. Seoul.
- Oloufa, A.A (2003) "Web-based tracking of school buses utilizing GPS and voice radio," in: *Proceedings of the 82nd Transportation Research Board Conference*. Washington, DC.
- Peters, J.I. and F.J. Mammano (1995) "Results of the Travtek system evaluation," in: *Proceedings of the World Congress on Applications of Transport and Telematics and Intelligent-Vehicle Highway Systems: Towards and Intelligent Transport System*, Vol. 2. Boston. Boston: Artech House.
- Polydoropoulou, A., D.A. Gopinath and M. Ben-Akiva (1997) "Willingness to pay for advanced traveler information systems: SmartTraveler case study," *Transportation Research Record*, 1588:1-9.
- Progressive (2002) *Progressive awarded second patent for usage-based auto insurance rating system*. Mayfield Village: Progressive ([http://www.progressive.com.newsroom/2nd\\_patent.asp](http://www.progressive.com.newsroom/2nd_patent.asp)).
- Rabah, M. and H. Mahmassani (2002) "Impact of information and communication technologies on logistics and freight transportation: example of vendor managed inventories," *Transportation Research Record*, 1790:10-19.
- Royal Automobile Club (2003) *Smartnav*. Bourne End: RAC ([http://www.rac.co.uk/travelservices/raclive/in\\_car\\_alerts](http://www.rac.co.uk/travelservices/raclive/in_car_alerts)).
- Stradling, S. (2001) "Intelligent speed adaptation: who wants it? Not those that need it!" *Traffic Engineering and Control*, 42:138-139.
- Sunberg, J. (2000) "Speed management: the need for an intelligent solution," *Traffic Technology International*, Feb./March.
- Tegaron (2003) *Tegaron Scout*. Bonn : T-Mobile Traffic ([http://www.tegaron.de/tegaron\\_en/index2.html](http://www.tegaron.de/tegaron_en/index2.html)).
- US Department of Transportation (2002) *Freight Information Real-Time System for Transport (FIRST)*. *Freight News*. Washington, DC: US Federal Highway Administration, Office of Freight Management and Operation.
- Victoria Transport Policy Institute (2002) *Pay-as-you-drive vehicle insurance*. Victoria: VTPI (<http://www.vtpi.org/tdm/79.html>).
- VisionAute (2003) [http://www.visionaute.tm.fr/visionaute\\_eng.html](http://www.visionaute.tm.fr/visionaute_eng.html).
- Ygnace, J., C. Benguigui, V. Delannoy, J. Remy, P. Auclair, J. Bosseboeuf, N. Schwab and V. Da Fonseca (2000) *Travel time/speed estimates on the French Rhone Corridor Network using cellular phones as probes. Final report for SERTI V program STRIP (System for Traffic Information and Positioning) project*. Lyon: INRETS.

*Chapter 26*

## TRAFFIC MONITORING USING GPS

CESAR QUIROGA

*Texas A&M University, San Antonio, TX*

### 1. Introduction

Traffic congestion has become a critical problem worldwide in many urban areas. For example, consider the USA. According to Schrank and Lomax (2002), 65% of the peak period person-kilometers traveled on US freeways in 2000 were under congested conditions. By comparison, the same percentage was 30% in 1982, and 52% in 1990. The corresponding percentages for principal US arterial roads were very similar. The total cost of congestion in 75 major urban areas in the USA – only including the effects of wasted time and fuel consumption – was about US \$68 billion, or US \$1160 per road traveler, in 2000.

Several performance measures are available to quantify congestion. A US example is the *Highway Capacity Manual* (HCM), which measures the level of service and volume/capacity ratio, and travel-time-based measures such as travel time, delay, speed, and queue duration (Transportation Research Board, 2000). The HCM measures are well understood by US transportation professionals but not by the traveling public. HCM measures are not well suited for multimodal comparisons. In some cases, they require the use of complex models that tend to be rather limited in their functionality. Finally, the HCM measures are difficult to use for long-term comparisons because the operational definition of concepts such as capacity tends to change over the years. In contrast, travel-time-based measures are easy to understand by both transportation professionals and the traveling public. They are flexible enough to describe traffic conditions at various spatial resolution levels – from specific locations to entire corridors – and temporal resolution levels – from minutes to years or decades. Travel-time-based measures translate easily into other measures (e.g. user costs), and can be used to validate travel-demand forecasting models (Laird, 1996). Travel-time-based measures are applicable across modes. All these reasons make travel-time-based measures extremely powerful, versatile, and desirable.

The number of transportation facilities, mostly freeways and major arterial roads, that are being retrofitted with roadside intelligent transportation system

(ITS) components is quickly growing in a number countries, such as the USA. These systems enable the capture of real-time traffic data that can be converted into near real-time travel time and speed data. Increasingly, ITS data are also being used as input to the transportation-planning process and as input to help optimize the internal operation of traffic management centers. However, most corridors in a typical transportation system are not equipped with roadside ITS components. The question then becomes how to collect travel time and speed data on those corridors in an efficient, cost-effective manner.

A number of vehicle-based techniques are available to collect travel time and speed data, including the traditional stopwatch-and-clipboard technique, the distance measuring instrument (DMI) technique, and the Global Positioning System (GPS) technique. With the stopwatch-and-clipboard technique, two technicians are required in the vehicle: one of them to drive the vehicle, and the other one to manually record the location and time of individual checkpoints, as well as length of time spent in queues. The level of accuracy obtained with this technique varies from technician to technician. Realistically, technicians can log checkpoint data if the distance between contiguous checkpoints is at least 0.5 km (Benz and Ogden, 1996). With this level of resolution, only average speeds can be calculated. At any level of resolution, queue information is highly subjective, because it depends on the technician's estimation of queue lengths. In addition to this, problems such as missing checkpoints or inaccurately marked checkpoints are common.

With the DMI technique, a device is connected to wiring from the vehicle's transmission to automatically record cumulative distances, time stamps, and speeds. These DMIs enable the production of more detailed distance-time profiles. When using DMIs, only one technician is needed in the vehicle. In some models, it is possible to store route and checkpoint location in the memory of the device to reduce some of the problems associated with missing or incorrectly marked checkpoints. However, DMIs are not free of difficulties. For example, Benz and Ogden (1996) reported a need for frequent DMI-vehicle unit calibration and tire pressure verification – an incorrect tire pressure can result in incorrect speed and/or distance readings. In addition, the accuracy of the survey still depends on the technician's ability to accurately mark the location of major checkpoints along the route, including the beginning and ending points. These characteristics limit the potential for using DMIs for long-term comparisons and geographic information system (GIS)-based applications, where the need to control the positional accuracy of individual measurements is important.

With the GPS technique, a GPS receiver is carried onboard to automatically record positions (in latitude and longitude), time stamps, and speeds at regular time intervals, e.g. every 1 s. Like the DMI technique, only one technician is needed in the vehicle. Unlike the DMI technique, the GPS technique is vehicle-independent and, in many ways, technician-independent. These characteristics, in

addition to the compatibility of the resulting GPS data with GIS applications and the decreasing cost and availability of GPS receivers, are helping to popularize the GPS approach for collecting travel time data. However, GPS receivers cannot determine the routes being surveyed or the distances traversed along those routes. As a result, procedures for determining routes and distances traveled – linear referencing – and for storing and aggregating the GPS data become essential. This chapter describes those procedures.

Recent research efforts have attempted to develop remote sensing, i.e. non-vehicle-based techniques to collect travel time and speed data. Examples include cell phone signal triangulation and aerial photography. While promising in some cases, results have been mixed for the most part, both in terms of accuracy and coverage, with high cost remaining one of the big challenges for implementing those techniques in practice.

## 2. Measuring travel times, speeds, and delays using GPS

Researchers and practitioners have attempted a number of approaches for deriving travel time, speed, and delay information from GPS data (Guo and Poling, 1995; Quiroga and Bullock, 1999a; Taylor et al., 2000; Faghri and Hamad, 2002; Quiroga and Perez, 2002). As an illustration, Figure 1 shows a generic workflow to calculate travel times, speeds, and delays using a GPS approach. Conceptually, the process involves building route files, mapping the GPS data to those routes, and, depending on the type of analysis required, calculating segment travel times and speeds and/or calculating intersection delays.

### 2.1. Generating routes, checkpoints, and segments

Unless the GPS-based travel time data collection is a one-time exercise (in which case, it is probably more efficient to just display the GPS data on a digital map and derive all travel time data manually), the first step is usually to generate routes, checkpoints, and segments in a GIS environment. While the process is initially more involved, using a GIS approach has three clear advantages:

- consistency – the same checkpoint and segment data can be used for all travel time runs, making travel times, speeds, and delays obtained from different travel time runs comparable;
- accuracy – using a digital map, recent aerial photography or even GPS data can result in a more accurate depiction of checkpoint locations than when trying to identify checkpoint locations while driving at highway speeds collecting travel time data;
- one time effort.

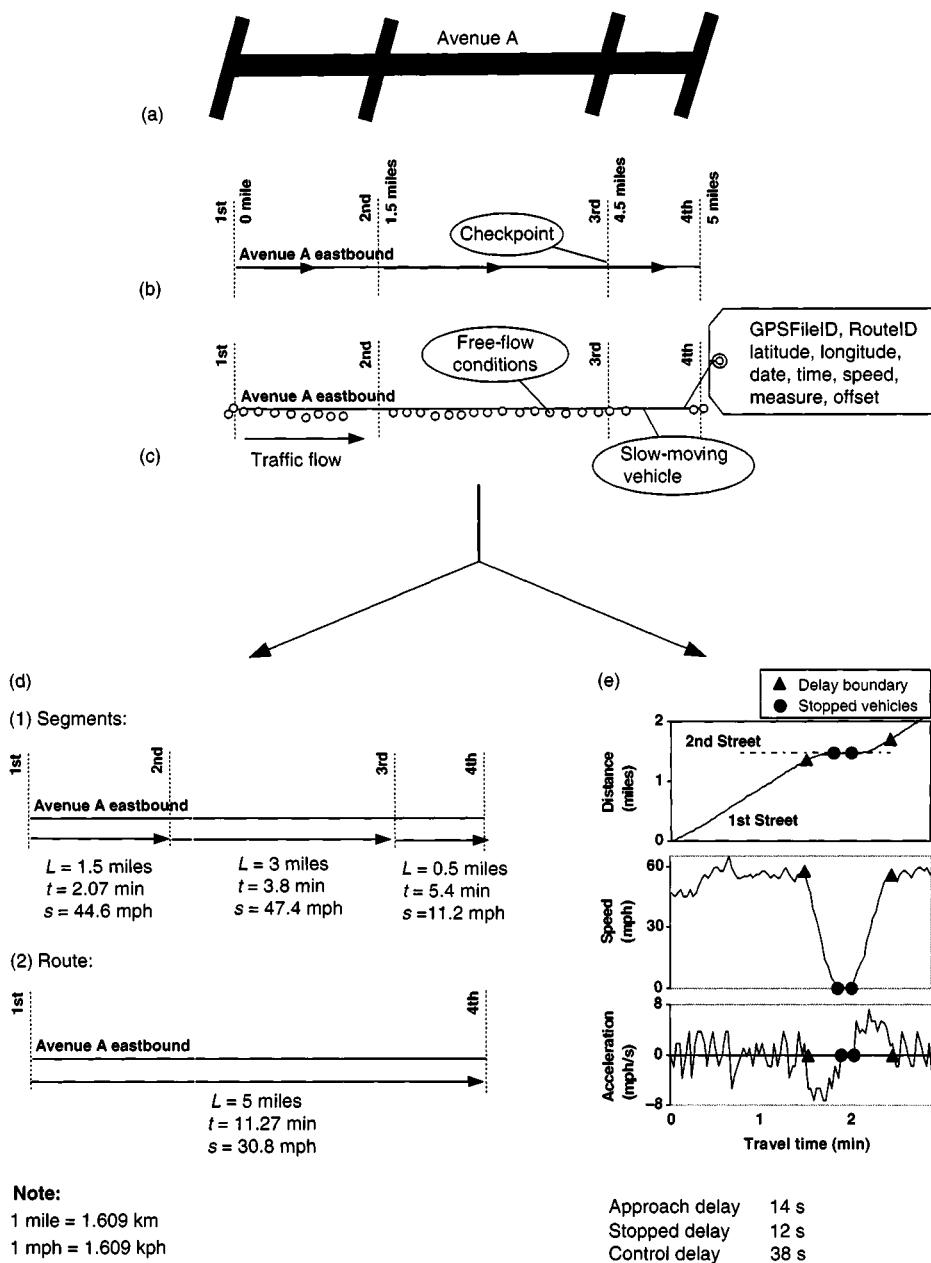


Figure 1. Procedure to calculate travel times, speeds, and delays. (a) City street map. (b) Electronic route file. (c) Linearly referenced GPS points. (d) Travel times and speeds. (e) Intersection delay.

Route files and checkpoint data can be generated using a variety of data sources, including digital city street maps (Figures 1a and 1b), recent geo-referenced aerial photography, and GPS data. Routes are directional elements that contain distance measures for all the vertices that characterize the route geometry. For example, the measure associated with the vertex located at the beginning of the Avenue A eastbound route on 1st Street is zero. Likewise, the measure associated with the vertex located at the end of the route on 4th Street is 8 km (5 miles). Readers should note that every vertex along the route is given a distance measure and that not necessarily every crossing street has a vertex. For example, if Avenue A follows a straight alignment between 1st Street and 4th Street, the route element will only have two vertices: one at 1st Street and the other one at 4th Street. Notice that each direction of travel corresponds to a different route. For example, Avenue A eastbound in Figure 1b represents the route used by eastbound vehicles. A second, westbound, Avenue A route would be used to represent the route used by westbound vehicles.

To calculate travel times, speeds, and delays along the route, it is necessary to identify checkpoints. Examples of checkpoints are signalized intersections, major unsignalized intersections, posted speed limit change locations, and school zone limits. For example, in Figure 1 there are four checkpoints, one for each crossing street. Each checkpoint has a distance measure along the route. For example, the distance measure associated with the checkpoint at 1st Street is zero. Likewise, the distance measure associated with the checkpoint at 2nd Street is 2.41 km (1.5 mile). Checkpoint distance measures can be calculated using GIS tools, which reduces the need for data collection personnel in the field to press buttons or record time stamps whenever they pass checkpoints while driving.

## 2.2. Linearly referencing GPS data

After generating routes, all GPS data collected during travel time runs can be mapped and linearly referenced to the route files. Linear referencing involves calculating a distance measure and an offset for every GPS point along the route (Figure 1c). The distance measure is an indication of the location of the GPS point along the route of interest, e.g. 2.56 km (1.59 miles), while the offset is an indication of the perpendicular displacement of the GPS point with respect to the route element alignment, e.g. 1.06 m. The linear referencing process is complete when the distance measures and offsets are added to the latitude, longitude, date, time, and speed data associated with each GPS point.

Standard GIS linear referencing tools should be used to map and linearly reference GPS data to routes. This way, the maximum possible error in the calculation of linear measures for individual GPS points is given by the positional accuracy of each GPS point (assuming, obviously, that the route was properly

generated). Because of the positional error associated with GPS data, calculating Euclidean distances between adjacent GPS points to measure distances along the route would result in errors that are substantially larger than the positional accuracy of individual GPS points, in effect degrading the accuracy of the linear referencing process.

### 2.3. Calculating segment travel times, speeds, and delays

Once the GPS data have been linearly referenced, it is possible to calculate segment travel times and speeds. Conceptually, the process involves time interpolating GPS locations to estimate the time stamps associated with checkpoints (Figure 2). These values are then used to calculate segment travel times and speeds (Figure 1d). For example, for the segment between 1st Street and 2nd Street (2.41 km (1.5 miles)), Figure 1d shows the travel time was 2.07 min, and the corresponding travel speed was 71.8 kph (44.6 mph). Adding the travel times for the three segments between 1st Street and 4th Street results in a total route travel time of 11.27 min, and a corresponding travel speed of 49.6 kph (30.8 mph).

After calculating segment travel time and speed data, it is possible to generate tabular and graphical reports. In the general case, there may be several travel time runs, and it is of interest to calculate average travel times and speeds for each segment and for the entire route. As Figure 3 shows, the number of travel time and speed data records per segment may be different. This could happen, for example, due to gaps in the GPS data, route overlapping, or complex route scheduling patterns. In Figure 3,  $t_{L_j}$  is the  $j$ th travel time record associated with segment  $i$ ,  $u_{ij} = L_i/t_{L_j}$  is the  $j$ th speed record associated with segment  $i$ , and  $L_i$  is the length of segment  $i$ .

Assume a representative segment travel time is given by the arithmetic average of all travel time values associated with a segment. Following Quiroga and Bullock (1999a), the total representative travel time and speed over all segments can be expressed as

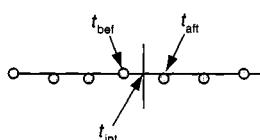


Figure 2. Time stamp interpolation.

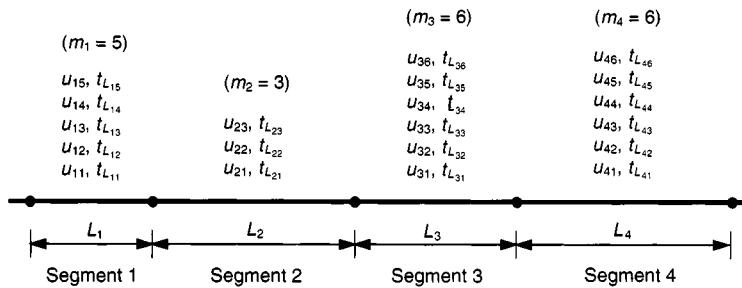


Figure 3. Sample segment speeds and travel times.

$$t_{T_L} = \sum_{i=1}^n \bar{t}_i = \sum_{i=1}^n \left( \frac{1}{m_i} \sum_{j=1}^{m_i} t_{L_{ij}} \right) = \sum_{i=1}^n \left( \frac{L_i}{m_i} \sum_{j=1}^{m_i} \frac{1}{u_{ij}} \right), \quad (1)$$

$$\bar{u}_L = \frac{L_T}{t_{T_L}} = \frac{\sum_{i=1}^n L_i}{\sum_{i=1}^n \bar{t}_i} = \frac{1}{\sum_{i=1}^n \left( (1/L_T m_i) \sum_{j=1}^{m_i} t_{L_{ij}} \right)} = \frac{1}{\sum_{i=1}^n \left( (L_i/L_T)(1/m_i) \sum_{j=1}^{m_i} 1/u_{ij} \right)}, \quad (2)$$

where  $t_{T_L}$  is the total representative travel time over all segments,  $n$  is the number of contiguous segments,  $\bar{t}_i$  is the average (arithmetic mean) of all travel time values associated with segment  $i$ ,  $m_i$  is the number of travel time and speed records per segment,  $L_T$  is the total length considered, and  $\bar{u}_L$  is the overall average speed.

Equations (1) and (2) are generic, and can be used to obtain average travel times and speeds for one or more segments and for one or more travel time runs per segment. Equation (2) represents a “weighted” harmonic mean of segment speeds, where the weight is the ratio of the length of each segment to the total length considered. One disadvantage of this equation is its sensitivity to outlying low speeds (which tend to occur in atypically adverse traffic conditions), resulting sometimes in very small average speeds. A more robust estimator of central tendency can be obtained by using median segment travel times instead of arithmetic mean segment travel times. The median speed formulation is

$$\bar{u}_L = \frac{L_T}{\sum_{i=1}^n t_{m_i}} = \frac{1}{\sum_{i=1}^n (t_{m_i}/L_T)} = \frac{1}{\sum_{i=1}^n [(L_i/L_T)(1/u_{m_i})]}, \quad (3)$$

where  $t_{m_i}$  is the median travel time associated with segment  $i$  and  $u_{m_i}$  is the median speed associated with segment  $i$ .

Quiroga (1997) provides a detailed comparison between equations (2) and (3) based on 26 000 segment speed records in Baton Rouge, Louisiana, USA.

Equations (1)–(3) assume an adequate number of travel time runs is available. Two types of formulations are normally used in practice for estimating required sample sizes: (1) based on sample ranges and (2) based on sample standard deviations. As an illustration, the formulation based on sample standard deviation is

$$n = \left( \frac{t_{\alpha} s}{\varepsilon} \right)^2, \quad (4)$$

where  $n$  is the required sample size,  $t_{\alpha}$  is the  $t$  distribution statistic for a confidence level of  $1 - \alpha$ ,  $s$  is the sample standard deviation, and  $\varepsilon$  is the user-selected allowable error or interval half-length.

Turner et al. (1998) and Quiroga and Bullock (1998) provide guidelines for application of the formulations, as well as the selection of  $\alpha$  and  $\varepsilon$ .

With segment travel times and speeds (either at the segment level or at the route level), additional measures can be derived (Lomax et al., 1997; Quiroga and Bullock, 1999a). For example, travel rate ( $r_L$ ) is the inverse of the segment speed, and is usually expressed in minutes per kilometer. While not readily understood by all audiences, travel rate provides a useful measure that can be averaged for a facility, geographic area, or mode. It can also be used to compare performance among transportation facilities more effectively than speed. Travel rate can be expressed as

$$r_L = t_L / L. \quad (5)$$

Delay ( $d_L$ ) is the difference between travel time and an acceptable travel time ( $t_{L_0}$ ) on a road segment. The acceptable travel time could be the travel time associated with the free flow speed or any other target value(s) as defined by local characteristics, e.g. one value during peak periods and a different value during off-peak periods. Delay is expressed as

$$d_L = t_L - t_{L_0}. \quad (6)$$

Total delay ( $D_L$ ) is the sum of delays for all vehicles traversing the segment during the time period (e.g. 15 min) for which travel time data are available, and is normally expressed in vehicle-minutes or vehicle-hours. Total delay can be expressed as

$$D_L = V d_L, \quad (7)$$

where  $V$  is the number of vehicles traversing the segment during the time period for which travel time data are available. Traffic volumes are established items in many roadway inventory databases.

Delay rate ( $d_{r_L}$ ) is the difference between the actual travel rate and an acceptable travel rate (e.g. 1.2 min/km during the peak period for an arterial corridor). Delay rate can be expressed as

$$d_{r_L} = \frac{t_L - t_{L_0}}{L} = \frac{d_L}{L}. \quad (8)$$

Relative delay rate ( $d_{r_R}$ ) is a dimensionless index that can be used to compare congestion on facilities, modes, or systems in relation to different mobility standards. It is calculated as the delay rate divided by the acceptable travel rate. Relative delay rate can be expressed as

$$d_{r_R} = \frac{t_L - t_{L_0}}{t_{L_0}} = \frac{d_L}{t_{L_0}}. \quad (9)$$

#### 2.4. Calculating intersection delays

Once the GPS data have been linearly referenced, it is possible to measure the delay experienced by the probe vehicle at all intersections on the route. Following Quiroga and Bullock (1999b), the procedure involves constructing distance–time profiles, speed–time profiles, and acceleration–time profiles to determine points where the vehicle decelerates, stops, resumes motion, and stops accelerating (Figure 1e). The corresponding time stamps and locations along the route are then used to provide a measure of control delay (Transportation Research Board, 2000), approach delay (the amount of control delay up to the intersection checkpoint), and stopped delay (the total length of time the vehicle traveled below a pre-specified crawling speed, say 3.2 kph).

Consider the distance–time diagram in Figure 4. Before point 1 on the distance–time diagram, the vehicle is moving at a uniform speed. From point 1 to point 2, the vehicle decelerates until it stops at point 2 to join the standing queue before the signalized intersection. The vehicle remains stopped between points 2 and 3. Between points 3 and 5, the vehicle accelerates until it reaches a uniform speed at point 5. Notice that when the vehicle crosses the stop bar (point 4), the vehicle is still accelerating.

In Figure 4, the stopped delay ( $d_s$ ) is given by

$$d_s = t_3 - t_2. \quad (10)$$

Similarly, the control delay ( $d_c$ ) is given by

$$d_c = (t_5 - t_1) - \frac{L_5 - L_1}{s_f}, \quad (11)$$

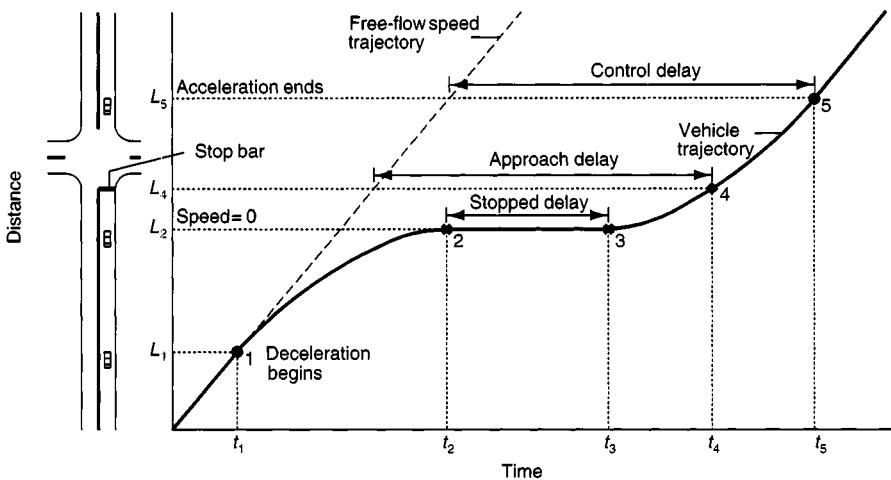


Figure 4. Schematic distance–time diagram depicting all delay terms at a signalized intersection.

where  $s_f$  is the free flow speed. For comparison purposes, Figure 4 also shows a generic approach delay  $d_{ap}$ , which is given by

$$d_{ap} = (t_4 - t_1) - \frac{L_4 - L_1}{s_f}. \quad (12)$$

Determining points 2, 3, and 4 in Figure 4 is straightforward. Determining points 1 and 5, however, is considerably more difficult. Quiroga and Bullock (1999b) have developed a procedure to determine points 1 and 5 by using information included in the speed–time and acceleration–time diagrams. The process involves using a “forward”-sweeping algorithm to detect when the vehicle acceleration becomes non-zero – i.e. when the vehicle starts to accelerate or decelerate – and a “backward” sweeping algorithm to detect when the vehicle’s acceleration becomes zero – i.e. when the vehicle stops accelerating or decelerating.

It is important to keep in mind that the delay being measured is the delay experienced by a single probe vehicle when it approaches a signalized intersection. As such, the delay measure might not be representative of the delay experienced by all the vehicles approaching that intersection. However, particularly in situations where the travel time study involves several runs on several intersecting corridors, the intersection delay experienced by the probe vehicles can be used to provide an area wide measure of delay for all the signalized intersections that were covered as part of the travel time study.

### 3. Data management

Regardless of the data collection technique used to collect travel time data, the data management component is critical. In general, that component should be built using a geographic relational or object data model and provide all the necessary interfaces and procedures to derive travel time, speed, and delay data accurately, reliably, and efficiently. Obviously, the structure of the data management component depends on the data collection technique used and the performance measures chosen.

As an illustration, this section describes software developed recently at the Texas Transportation Institute (Quiroga and Perez, 2002). This application, called GPS-based Evaluation of Travel Time (GETT), is a data-reduction and data-reporting tool designed to assist engineers and planners during the execution of travel time studies. GETT offers the capability to play back GPS data files, linearly reference GPS data to routes on the highway network, calculate travel times and speeds at various levels of spatial and temporal resolution, calculate intersection delays, and generate reports. GETT also offers the capability to store all data in a relational database environment for subsequent retrieval and analysis, and includes several filters and procedures to assist in the data quality assessment/control process.

#### 3.1. Architecture

GETT was developed as a standalone tool using Microsoft Visual Basic and Environmental Systems Research Institute (ESRI) MapObjects components (Figure 5). In the current version, GETT can open and display GPS data files, as well as ESRI ArcView 3.2 shape files containing point, polyline, or polylineM features. Point and polyline shape files can be used to provide map background information. PolylineM shape files are used to represent routes based on which GETT linearly references GPS data.

The database (in Microsoft Access 2000 format) follows a relational database model (Figure 6). The table "GPSFiles" contains an index of GPS files generated in the field, including starting and ending dates and times, universal coordinated time (UTC) offset, number of records, and a reference to a unique status ID in the table "GPSStates." The table "GPSPoints" contains a repository of all GPS data collected in the field. Each record contains date, time, latitude, longitude, speed, route ID, measure, and offset. The table "Routes" contains a listing of all routes found in the network. The table "Segments" contains a listing of route segments. In general, a route is composed of a series of adjacent segments. GETT allows users to store and use several segmentation tables, each one containing a separate segmentation configuration for any route in the network. The table

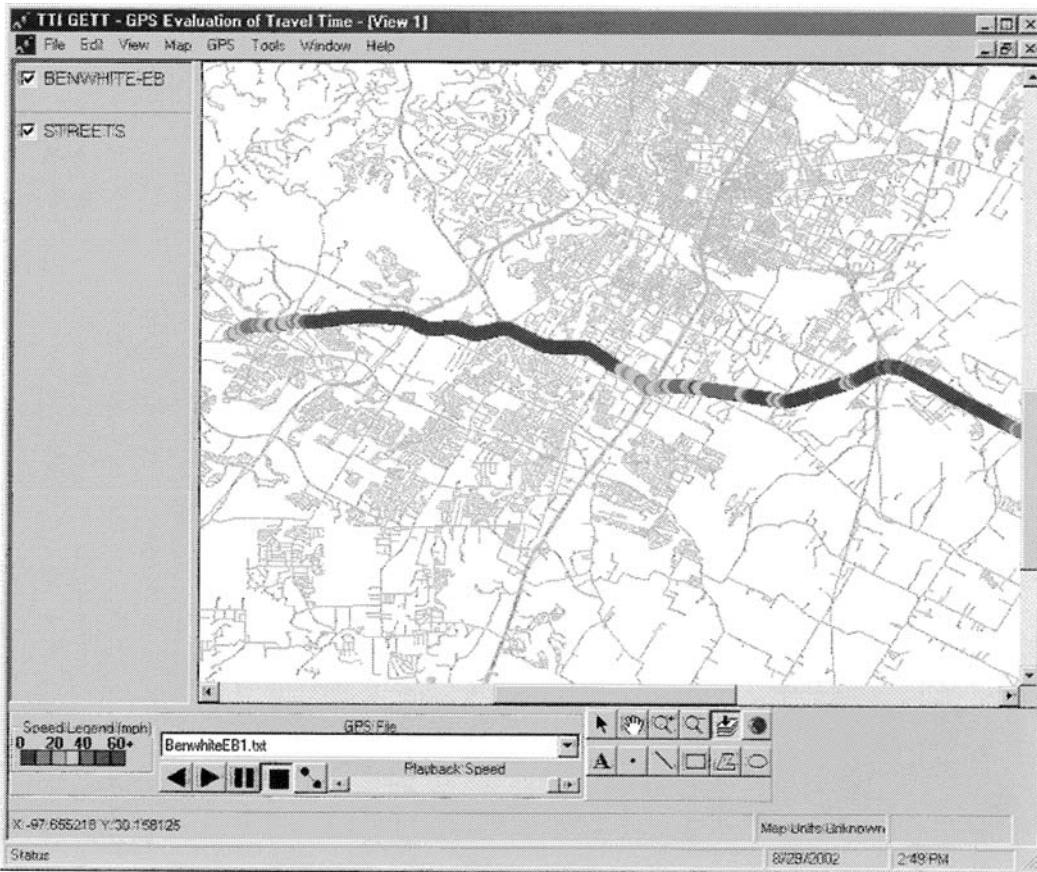


Figure 5. The GETT user interface.

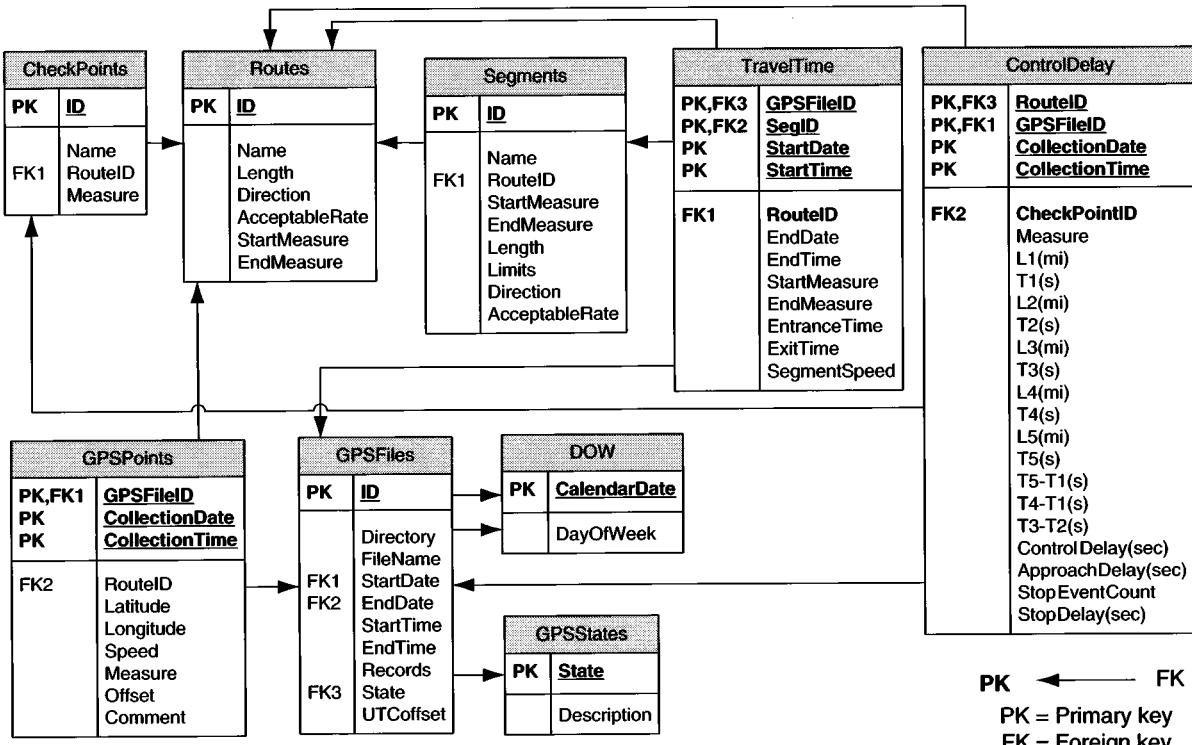


Figure 6. Database schema.

“TravelTime” contains segment travel time and speed data that result from the aggregation of linearly referenced data using the segmentation scheme defined by the table “Segments.” The table “CheckPoints” contains a listing of checkpoints, route ID references, and corresponding distance measures along the routes. This table is used for the calculation of intersection delays. Although, in general, the contents in the table “CheckPoints” are consistent with those in the table “Segments,” users can define separate checkpoint data tables depending on project needs. The table “ControlDelay” contains all of the data elements needed to calculate intersection delay, approach delay, and stopped delay.

### *3.2. Linear referencing and computation of travel time*

GETT uses a two-step approach for the computation of travel times and speeds (Figure 7):

- Linearly reference GPS data. Linear referencing means calculating a cumulative distance along a specific route for each GPS point. GETT uses polylineM shape files that represent routes. PolylineM shape files contain information about routes and cumulative distances for all vertices along the linear features. In the current version, an ArcView script is used to generate route files.
- Calculate segment travel times and speeds. GETT uses a user-defined segmentation table to convert point linearly referenced GPS data into segment travel times and speeds. GETT uses a time interpolation scheme to calculate time stamps at the checkpoints defined by the segmentation table.

With the segment travel time and speed data, it is possible to generate tabular and graphical reports. Two types of tabular summary report are possible in the current version of GETT: segment reports and route reports. Segment summary reports document average travel times, speeds, and delays at the segment level. Route summary reports document average travel times, speeds, and delays at the route level. GETT calculates delays with respect to a reference (or desired) travel rate value. GETT also calculates a delay ratio that represents the relative amount of delay rate with respect to the average travel rate.

In addition, GETT provides the capability to export data and reports for additional post-processing. For example, with the segment travel time summary table (which is stored in Access format) it is possible to generate color-coded maps to document congestion problems in ArcView format (Figure 8).

GPSFileID	CollectionDate	CollectionTime	Latitude	Longitude	Speed	RouteID	Measure	Offset
1	4/8/2002	7:57:41 AM	30.23068277	-97.87945581	29.028	benwhite	0.5560012	3.0813596E-06
1	4/8/2002	7:57:42 AM	30.23074938	-97.87934354	29.405	benwhite	0.5541314	2.7497338E-06
1	4/8/2002	7:57:43 AM	30.23081512	-97.87923286	28.45	benwhite	0.5721567	3.8130133E-06
1	4/8/2002	7:57:44 AM	30.23088117	-97.87912191	29.243	benwhite	0.58022	6.0598804E-06
1	4/8/2002	7:57:45 AM	30.23095165	-97.87900487	31.879	benwhite	0.5887525	7.7460281E-06
1	4/8/2002	7:57:46 AM	30.23102757	-97.87987881	34.233	benwhite	0.5979428	9.5550904E-06
1	4/8/2002	7:57:47 AM	30.23110711	-97.87874356	36.341	benwhite	0.6077401	1.3123983E-05
1	4/8/2002	7:57:48 AM	30.23119157	-97.87860011	38.318	benwhite	0.6181346	1.6826258E-05
1	4/8/2002	7:57:49 AM	30.23126013	-97.87845041	39.697	benwhite	0.6289958	2.0332616E-05
1	4/8/2002	7:57:50 AM	30.23137112	-97.87829473	41.167	benwhite	0.6402544	2.4920817E-05

(a)

GPSFileID	SegID	RouteID	StartDate	EndDate	StartTime	EndTime	StartMeasure	EndMeasure	EntranceTime	ExitTime	SegmentSpeed
1	289	benwhite	4/8/2002	4/8/2002	7:57:41 AM	7:58:24 AM	0.5560012	0.9438663	7:57:29 AM	7:58:24 AM	31.8838
1	290	benwhite	4/8/2002	4/8/2002	7:58:25 AM	8:00:53 AM	0.9480827	1.378961	7:58:24 AM	8:00:53 AM	10.51332
1	291	benwhite	4/8/2002	4/8/2002	8:00:54 AM	8:02:55 AM	1.380311	2.456887	8:00:53 AM	8:02:55 AM	31.96515
1	292	benwhite	4/8/2002	4/8/2002	8:02:56 AM	8:03:38 AM	2.475418	3.305934	8:02:55 AM	8:03:39 AM	70.07231
1	293	benwhite	4/8/2002	4/8/2002	8:03:39 AM	8:04:04 AM	3.32461	3.775327	8:03:39 AM	8:04:05 AM	64.6791
1	294	benwhite	4/8/2002	4/8/2002	8:04:05 AM	8:04:21 AM	3.794513	4.103167	8:04:05 AM	8:04:21 AM	73.60143
1	295	benwhite	4/8/2002	4/8/2002	8:04:22 AM	8:05:00 AM	4.123712	4.871794	8:04:21 AM	8:05:00 AM	70.68985
1	296	benwhite	4/8/2002	4/8/2002	8:05:01 AM	8:05:09 AM	4.890858	5.046448	8:05:00 AM	8:05:10 AM	65.22755
1	297	benwhite	4/8/2002	4/8/2002	8:05:10 AM	8:05:21 AM	5.065415	5.269928	8:05:10 AM	8:05:21 AM	69.88148
1	298	benwhite	4/8/2002	4/8/2002	8:05:22 AM	8:05:29 AM	5.289069	5.429969	8:05:21 AM	8:05:30 AM	68.8673
1	299	benwhite	4/8/2002	4/8/2002	8:05:30 AM	8:05:41 AM	5.450395	5.669644	8:05:30 AM	8:05:41 AM	76.14867
1	300	benwhite	4/8/2002	4/8/2002	8:05:42 AM	8:06:03 AM	5.688939	6.096576	8:05:41 AM	8:06:04 AM	67.49195
1	301	benwhite	4/8/2002	4/8/2002	8:06:04 AM	8:06:16 AM	6.116528	6.355836	8:06:04 AM	8:06:16 AM	76.66826
1	302	benwhite	4/8/2002	4/8/2002	8:06:17 AM	8:06:40 AM	6.37604	6.825893	8:06:16 AM	8:06:41 AM	69.43277

(b)

Figure 7. Sample records in the tables (a) “GPSPoints” (linearly referenced GPS data) and (b) “TravelTime” segment travel time and speed data.

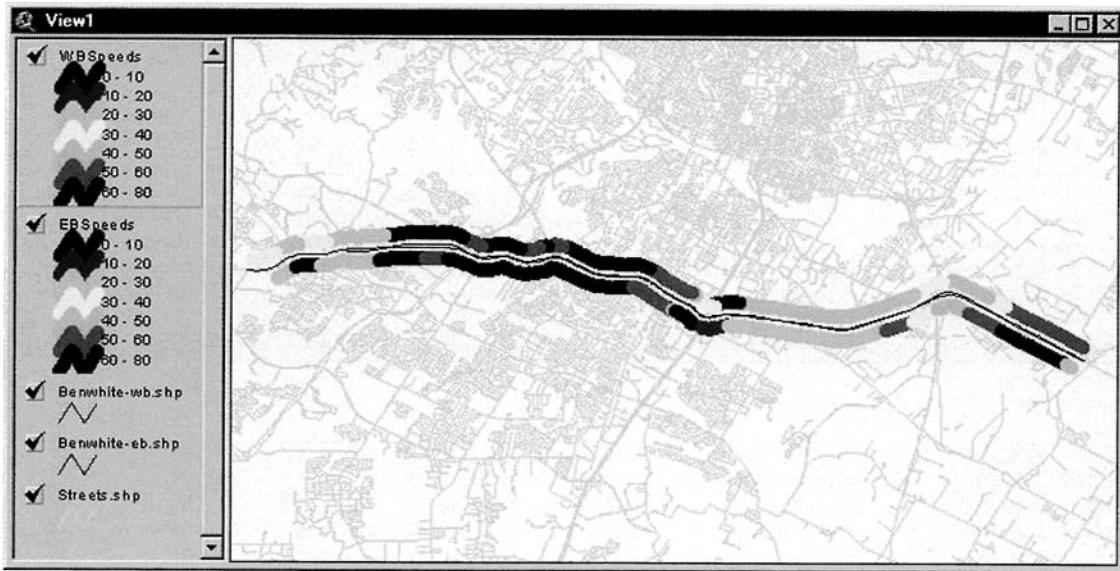


Figure 8. Segment speed map (the route reproduced here in shades of gray is color coded in the original screenshot).

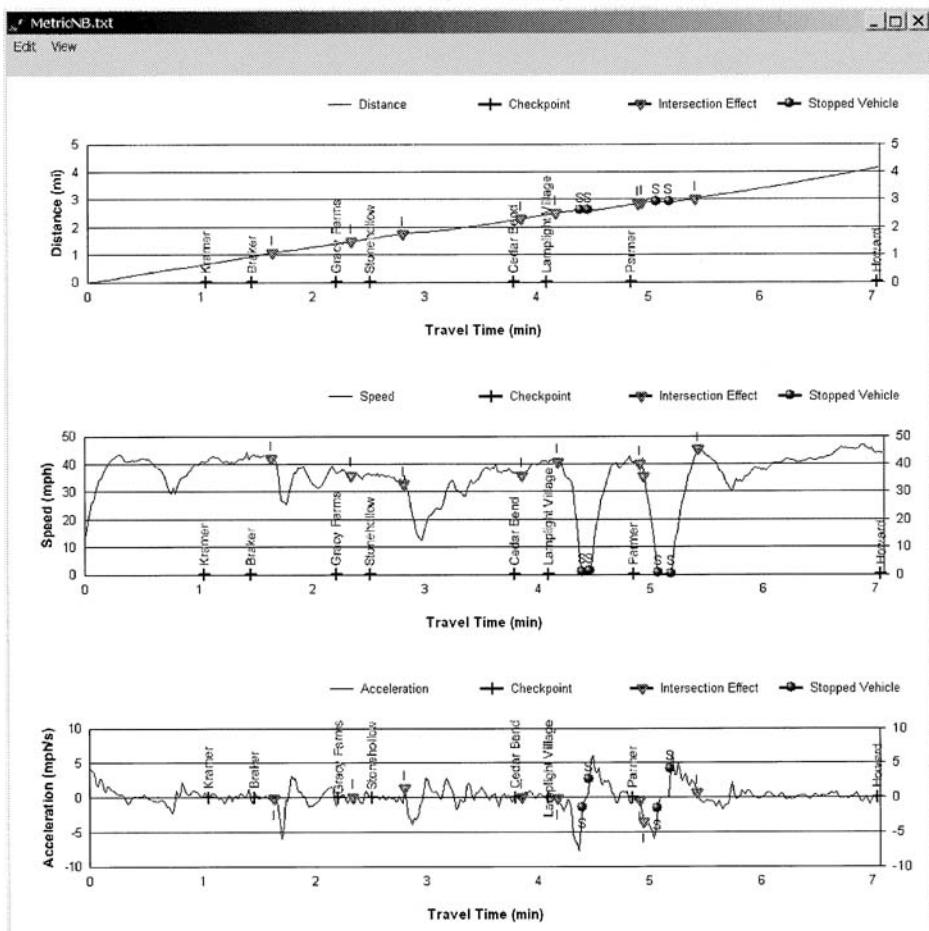


Figure 9. Distance, speed, and acceleration diagrams showing control delay and stop delay points.

### 3.3. Intersection delay

GETT can be used to calculate delay (control delay, approach delay, and stop delay) at critical intersections along routes. GETT uses “forward”- and “backward”-sweeping algorithms to detect the beginning and ending of major acceleration and deceleration events. The algorithms filter out minor acceleration “noise” that occurs during the normal driving process and focus on the detection of points needed for the calculation of deceleration delay, stop delay, and acceleration delay. GETT also provides users with the capability to interactively refine the location of delay control points (Figure 9). Once the

The screenshot shows a Windows application window titled "Intersection Delay - Detailed Report". The window has a standard title bar with icons for minimize, maximize, and close. Below the title bar is a toolbar with icons for zoom and orientation. The main area contains a table of data with 17 columns and 10 rows. The columns are labeled as follows: Checkpoint Name, Checkpoint Milepost, Checkpoint Time, L1 (mi), T1 (sec), L2 (mi), T2 (sec), L3 (mi), T3 (sec), L4 (mi), T4 (sec), L5 (mi), T5 (sec), T5-T1 (sec), T4-T1 (sec), T3-T2 (sec), Control Delay (sec), Approach Delay (sec), Stop Event Count, and Stop Delay (sec). The first row contains the column headers. Rows 2 through 10 contain data for four different locations: Gracy Farms, Cedar Bend, Parmer, and another location whose name is partially visible. The data includes GPS file numbers (3, 12, 24, 23) and various delay and control values.

Checkpoint Name	Checkpoint Milepost	Checkpoint Time	L1 (mi)	T1 (sec)	L2 (mi)	T2 (sec)	L3 (mi)	T3 (sec)	L4 (mi)	T4 (sec)	L5 (mi)	T5 (sec)	T5-T1 (sec)	T4-T1 (sec)	T3-T2 (sec)	Control Delay (sec)	Approach Delay (sec)	Stop Event Count	Stop Delay (sec)
Date:	Tue, Jun 20, 2000	GPSFile:	3																
Route:	metricOn																		
Gracy Farms	1.40	11:48:58 AM	1.07	99					1.40	133	1.48	141	42	34		12.46	10.43		
Cedar Bend	2.25	11:50:33 AM	1.75	169					2.24	228	2.28	232	63	59		24.80	23.72		
Parmer	2.81	11:51:36 AM	2.49	251	2.60	264	2.60	268	2.80	291	2.86	296	45	40	4	19.02	17.98	1	
																		4.00	
Pages:  1																			

Figure 10. Intersection delay detailed report. Note: the column headers follow the definitions given by eqs (9)–(11).

delay control points are determined, it is possible to generate tabular reports (Figure 10).

#### 4. Summary

Up-to-date congestion data are critical for evaluating improvement strategies that focus on mobility and system reliability such as corridor improvements, signal coordination efforts, and deployment of ITS components. This chapter described procedures for measuring travel times, speeds, and delays using a GPS approach.

The chapter has also summarized an application (GETT) developed at the Texas Transportation Institute to automate the GPS data reduction and data reporting processes. GETT uses linearly referenced GPS data and checkpoints along the routes of interest to calculate partial and cumulative travel times, speeds, and delays. The application also measures three elements of intersection delay: control delay, approach delay, and stopped delay. The application produces detailed and summary tabular reports, and output data can be used to generate color-coded maps depicting congestion problems along corridors.

The chapter has focused on a basic description of procedures to obtain travel time, speed, and delay data using GPS receivers on board probe vehicles. Owing to space limitations, it was not possible to discuss in detail examples of specific applications of the procedures, nor was it possible to fully discuss issues such as the required sample size and accuracy issues. Readers who are interested in these topics could review some of the references cited here. The decision to focus on the description of the procedures was deliberate. GPS has become a commodity. Researchers and practitioners are now much more familiar with the technology, and new GPS applications are reported regularly for surveying, navigation, fleet management, household travel surveys, travel time studies, and construction automation, among other uses. Despite these advances, there is relatively little documentation available to transportation professionals about strategies on how to manage the large amounts of data that result from the operation of GPS receivers. This chapter is intended to help fill that vacuum.

#### References

- Benz, R.J. and M.A. Ogden (1996) "Development and benefits of computer-aided travel time data collection," *Transportation Research Record*, 1551:1-7.
- Faghri, A. and K. Hamad (2002) "Application of GPS in traffic management systems," *GPS Solutions*, 5:52-60.
- Guo, P. and A.D. Poling (1995) "Geographic information systems/global positioning systems design for network travel time study," *Transportation Research Record*, 1497:135-139.
- Laird, D. (1996) "Emerging issues in the use of GPS for travel time data collection," in: *Proceedings of the National Traffic Data Acquisition Conference*, Vol. 1. Albuquerque.

- Lomax, T.J., S. Turner, G. Shunk, H.S. Levinson, R.H. Pratt, P.N. Bay and G.B. Douglas (1997) *Quantifying congestion*, Final report. Washington, DC: National Cooperative Highway Research Program, Transportation Research Board.
- Quiroga, C.A. (1997) "An integrated GPS-GIS methodology for performing travel time studies," Ph.D. dissertation. Baton Rouge: Louisiana State University.
- Quiroga, C.A. and D. Bullock (1998) "Determination of sample sizes for travel time studies," *ITE Journal on the Web*, Aug.:92–98 (<http://www.ite.org>).
- Quiroga, C.A. and D. Bullock (1999a) "Travel time information using GPS and dynamic segmentation techniques," *Transportation Research Record*, 1660:48–57.
- Quiroga, C.A. and D. Bullock (1999b) "Measuring control delay at signalized intersections," *Journal of Transportation Engineering*, 125:271–280.
- Quiroga, C.A. and M. Perez (2002) *GETT (GPS-based Evaluation of Travel Time) 1.0 user manual*. Austin: Texas Transportation Institute, Texas Department of Transportation.
- Schrank, D.L. and T.J. Lomax (2002) *The 2002 urban mobility report*. College Station: Texas Transportation Institute.
- Taylor, M.A., J.E. Woolley and R. Zito (2000) "Integration of the Global Positioning System and geographical information systems for traffic congestion studies," *Transportation Research Part C*, 8:257–285.
- Transportation Research Board (2000) *Highway capacity manual*. Washington, DC: Transportation Research Board, National Research Council.
- Turner, S., W. Eisele, R. Benz and D. Holdener (1998) *Travel time data collection handbook*, Report FHWA-PL-98-035. Washington, DC: Federal Highway Administration.

*Chapter 27*

## OTHER TRANSPORTATION APPLICATIONS OF GPS

SHAUNA L. HALLMARK

*Iowa State University, Ames, IA*

### 1. Introduction

The widespread availability of Global Positioning System (GPS) devices in recent years has fundamentally altered the way spatial data are collected. A GPS receiver captures signals from an orbiting network of GPS satellites. Using information from at least four satellites, a GPS receiver is able to calculate the distance from each satellite and, through trilateration, calculate its position on earth. Consequently, a GPS can fairly accurately locate and report planar coordinates (usually latitude and longitude) for a particular point as well as altitude, date, time, speed, and heading. Corresponding attribute data can be either manually or digitally data-logged, and then linked to the point feature in a database. Line and polygon features may also be collected with GPS by connecting lines between two or more coordinate pairs. Transportation-related GPS applications are numerous, and range in complexity from simple applications, such as collection of bus stop locations, to sophisticated applications, such as real-time vehicle tracking for intelligent transportation systems (ITS).

GPS is used by numerous transportation agencies, including metropolitan planning organizations (MPOs), departments of transportation (DOTs), county and local government agencies, and transportation engineering consultants. GPS is increasingly being used for transportation applications by the private sector as well, with the advent of in-vehicle navigation systems and fleet-tracking systems.

The US Federal Highway Administration (2000) surveyed state and local DOTs about the use of GPS at their agencies. A total of 32 states responded. Figure 1 illustrates the most common uses of GPS by those agencies. The number reflects those who actually stated that they used GPS for a particular application, and may have included responses indicating that they are evaluating or implementing such an application. As shown, most agencies used GPS for surveying purposes. The next most common application was the use of GPS for collecting roadway inventory data. The data collected varied by state, but included items such as locations of signs, traffic signals, or mileposts. The use of GPS to delineate

wetland and archeological site boundaries and locate sensitive features was another common application. Using automatic vehicle location (AVL) to manage fleets and either correct or create new centerline information were also popular uses of the technology. Several GPS applications are discussed in more detail in the following sections.

## 2. Centerline mapping

Most transportation agencies, such as state DOTs, maintain some type of digital road database. Original street databases were often created by digitizing paper maps and are frequently characterized by a number of errors due to the scale of the original source map, digitizing error, etc. Noronha et al. (2000) identify two common types of error in cartography as spatial alignment errors and insufficient polyline resolution. Alignment errors include missing segments, inaccurate representations of street segments, and spatial inaccuracies. Road segments may be missing as a result of digitizing error, errors in the source data, or because the database is not current. Inaccurate representation of the street centerline is usually due to digitizing error, but may also occur when street alignment

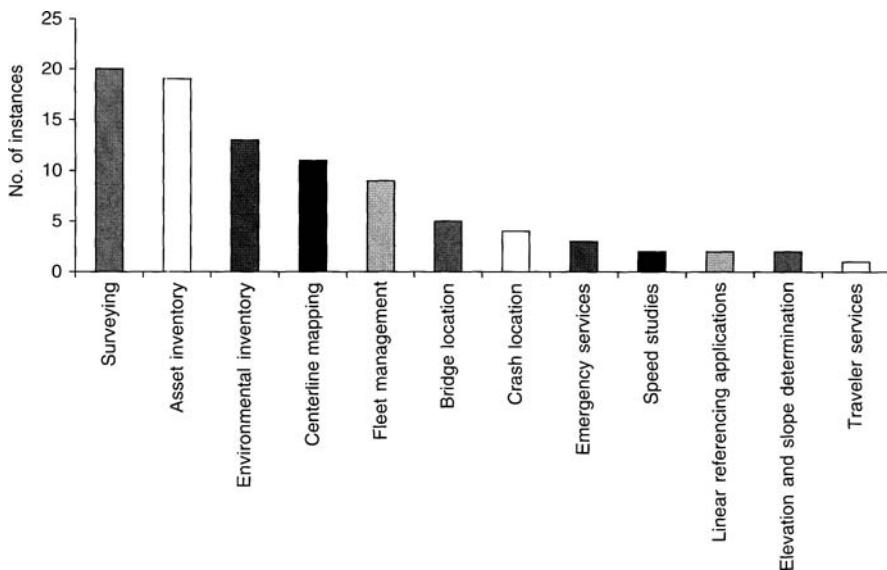


Figure 1. Use of GPS at state and local DOTs in the USA. (Data source: US Federal Highway Administration, 2000.)

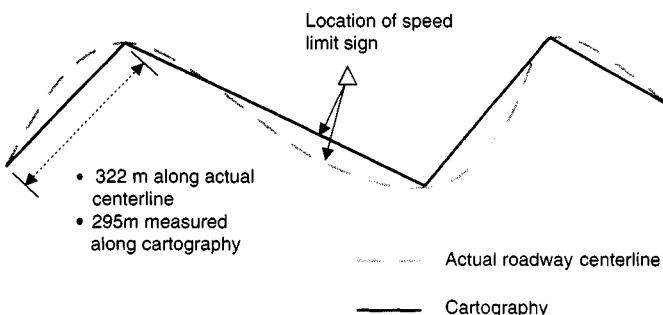


Figure 2. Insufficient polyline resolution.

changes but the database is not updated. Spatial accuracy is a measure of where cartography is located in relationship to actual location. A cartographic representation of a street centerline may correctly represent roadway geometry, but be offset from its actual position. Insufficient polyline resolution occurs when a line segment lacks “shape,” as shown in Figure 2. It is the result of either basing cartography on maps that lack spatial detail, digitizing an insufficient number of shape points along a line, or not using true spiral and circular curve descriptions of the road.

Although existing centerline databases may be adequate for many traditional applications such as pavement management, a number of emerging map-based applications require road centerlines that are much more accurate and current. Linear referencing methods (LRM) require centerlines of sufficient spatial accuracy to allow linear offset to be measured. Linear referencing locates objects in terms of their distance (offset) and direction along a segment from a known location. Centerline databases that are spatially accurate, correctly represent roadway geometry, and are temporally correct are critical for applications such as location based services (LBS) and vehicle navigation systems. GPS is one method that agencies are using to create, correct, verify, and update cartographic representations of roadway databases. Britch and Fitch (1997) describe the use of GPS to collect line data for Adopt-a-Highway segments in the state of Virginia. They compared line segment lengths collected using GPS to test segments, and found that most of the GPS line segments were within 1% of the length of the test segments. South Dakota began a project in 1995 to update county highway maps and record actual mileage for all roads in the state. The actual location of roads was mapped using GPS, including attributes such as road width, number of lanes, and mile reference markers. Other features such as the location of bridges and power substations, were recorded as well. One of the common applications of GPS at state and local agencies is to verify existing roadway databases, create new

databases, or update existing maps with the location of new facilities (US Federal Highway Administration, 2000). Masters et al. (1994) indicate that map scales of up to 1:10 000 can be created using differential GPS (DGPS).

### **3. Inventory management**

#### *3.1. General*

A number of inventory elements are collected and maintained by US transportation agencies to meet the data requirements of the Highway Performance Monitoring System or to support the numerous functions within agencies. GPS is frequently used to collect those items. Handheld GPS receivers are common tools to collect inventory items directly in the field. Poling et al. (1994) describe the use of GPS for collection of sign inventory data. The GPS system utilized in the study provided location accuracies from 2 m to 5 m with differential correction. Data were collected by personnel individually locating each sign. Sign attributes were also recorded. Poling et al. reported that manual methods had to be used in locations where tall buildings blocked satellite signals. Britch and Fitch (1997) describe the use of handheld mapping-grade GPS to collect locational and attribute data for the Virginia Department of Transportation. Several inventory elements, including highway–railroad grade crossings, were located using the devices. Ellis et al. (2001) describe the architecture for a prototype platform to collect the location of utility facilities located within the highway right-of-way to manage utility data to support a utility management system. The system uses GPS, GIS, and Internet technologies. Location is collected as well as the ownership, purpose, size, type, and shared and multiple uses.

A number of US states reported using GPS for some aspects of collection of inventory features. Of the 32 states surveyed by the US Federal Highway Administration (2000), 60% used GPS for inventory management. Some of the elements that state and local agencies collected included:

- signs;
- noise receptors;
- underground storage tanks;
- railroad crossings;
- reference posts;
- guard rails;
- median barriers;
- drainage structures;
- county boundaries;
- section corners;

- fire hydrants;
- utilities;
- bus stops, routes, signs, and shelters;
- light posts;
- traffic signals.

### 3.2. Mobile mapping systems

Mobile mapping systems (MMS) are more sophisticated systems than those described so far, and are used for the collection of roadway inventory features (Karimi et al., 2000). MMS are essentially moving platforms that have integrated multiple sensor/measurement systems for data collection along with data processing and management software that allows collection of spatial features. Vans are frequently used as the data collection vehicle. MMS are able to function with no or limited ground control using single or multiple GPS base stations. Most MMS consist of an imaging system, a geo-referencing system, and a photogrammetric measurement system (Burtch, 2003). The geo-referencing system includes a GPS and some form of dead-reckoning unit, such as an inertial navigation system (INS), which work in conjunction for position and navigation. Many systems also have distance-measuring instruments (DMIs) as well. The DMI and INS are used as back-ups when the GPS signal is lost. The DMI also records distance, and triggers data collection at specific intervals. The photogrammetric measurement system consists of several synchronized cameras, resulting in stereo images (Lambda Tech, 2003). Images are processed with the GPS/INS data, which provide coordinates, and rotational information that is linked to the images so that the position of an object can be extracted. An object is selected in two or more overlapping images, which are then processed using photogrammetric software. Attributes such as height and width can be determined as well as position. MMS are used in such applications as right-of-way (ROW) imaging; locating centerlines in GIS, sign and signal inventory; locating intersections, infrastructure inventory; utility pole mapping; and incident mapping (Kut and Baraniak, 2000). Digital cameras are also used to record visual images. One of the main advantages is that data collection can be accomplished while driving at normal highway speeds.

## 4. Automatic vehicle location

GPS provides the ability to track vehicles in real time, and is used in fleet management systems. With GPS, vehicles can be tracked and monitored, resulting in more efficient dispatch of vehicles to meet customer demand. More efficient

dispatch results in more optimal allocation of resources and shorter waiting times. This method also requires network information to establish the closest taxi. Other benefits include the following (Drane and Rizos, 1998):

- the location of cargo can be monitored when necessary;
- customers can be provided with the location of vehicles for pick-up or delivery;
- drivers can be provided with navigation assistance when negotiating unfamiliar areas.

Performance measures can be collected, reported, and analyzed, such as time in transit, total distance traveled, or average speed.

#### *4.1. In-vehicle navigation systems*

An increasingly common use of AVL is in-vehicle navigation systems. An in-vehicle navigation system consists of a GPS receiver to establish position, a gyroscope to measure movement, and a means of processing electronic signals from the speedometer to measure driving distances. Measurements for the three target variables are translated into the latitude, longitude, and direction of travel. Information is processed by the unit and overlain with a navigable map that can display a vehicle's location, calculate the distance between destinations and turns, and create a route (e.g. Navteq). Units vary by capability, but most units are capable of positional accuracy within 15 m and accuracy for velocity within 0.05 m/s. With DGPS, positional accuracies of 3–5 m are possible.

AVL can also be used to provide roadside assistance to drivers. Systems have come online recently to provide roadside assistance and emergency and directional information to motorists. RoadStar is typical of these services, and provides coverage in North America. The system consists of a cell phone, personal digital assistant (PDA) or notebook PC and a separate GPS accessory that attaches directly to the battery compartment. The GPS accessory locates the user's position, and transmits the data to a central call center. With the system, users can receive roadside assistance, trip planning, location-based directory assistance, and turn-by-turn navigation based on the user's targeted GPS location (Woods, 2002).

#### *4.2. Fleet management*

AVL is also used in management of fleet vehicles. Czerniak and Reilly (1998) discuss the use of this technology to track transit buses. Vehicle tracking is used for scheduling, to optimize route structure, and to monitor location and speed. The Hong Kong fire service department is moving toward using an automatic vehicle

locating system to dispatch fire and ambulance services (Advantec, 2002). The system was designed to track and update vehicle position, status, dispatch information, and other variables every 6 s. McLellan et al. (1993) describe NavTrax, which is a dispatch-type AVL and navigation system. The system uses a robust GPS-based system with an integrated dead-reckoning sensor and cellular or UHF radio-type technology for fleet management.

The US Federal Highway Administration (2002) reported numerous uses of GPS for fleet management. For example, fleet management systems were used to track:

- snowplows;
- vehicles carrying hazardous waste;
- emergency management vehicles;
- highway patrol vehicles;
- roadside assistance vehicles;
- transit vehicles;
- heavy trucks.

#### *4.3. Concept winter vehicle*

Winter maintenance, such as clearing ice and snow from roads, is one of the major expenses incurred by transportation agencies in areas that receive significant snowfall. Pilot studies have been conducted that equip snowplows with technology such as road temperature sensors, cameras, and GPS receivers: the temperature sensors help determine the best anti-icing treatments; the cameras monitor spreading; and GPS is used to automatically track the vehicles so that the agencies can ensure that resources, such as labor or de-icing or anti-icing materials, are efficiently utilized. Real-time road conditions can also be provided. Concept snowplows have been tested in Wisconsin, Michigan, and Iowa (*Public Works*, 2001; US Federal Highway Administration, 2002). Automatic tracking of fleet management of existing snowplows is also undertaken in several other US states (US Federal Highway Administration, 2000).

## **5. Safety**

### *5.1. Crash location*

Frequently, positional information for crash locations is based on estimates made by the emergency services, such as the police, in the field. Given limited time and tools to accurately position crashes, accidents are often located using rough

estimates or by linking them to locations that are easily identifiable, such as mileposts. As a result, crash location information is only as accurate as the reporting personnel's estimate (Miller and Karr, 1998). Without fairly accurate spatial locations, it is difficult to link roadway variables to accidents, and consequently difficult to select appropriate countermeasures. As a result, GPS is being used as a more accurate method to locate crashes. Graettinger et al. (2001) describe the evaluation of two GPS units combined with a GIS basemap to locate vehicle crashes for the state of Alabama. Currently the state locates crashes by estimating and hand recording. This method results in crashes often being placed at easily identifiable locations such as intersections, which can result in significant spatial inaccuracies. Consequently, it is difficult to pinpoint roadway variables that are hazardous. The study showed that even inexpensive GPS systems are capable of locating a vehicle crash to within a 12 km radius, allowing "hot spot" locations with multiple crashes to be identified. Miller and Karr (1998) also evaluated GPS for use in crash locations. They compared the accuracy of handheld GPS devices to conventional crash location methods. GPS can also be used to delineate emergency and fire service boundaries so that dispatchers can expedite dispatch of emergency vehicles.

## 5.2. *On-board crash notification systems*

Another use of GPS in safety applications is the use of an on-board crash notification system, which consists of an accelerometer and a GPS receiver. A sudden change in acceleration triggers the possibility of a crash. Position is determined by the GPS device, and the system then notifies emergency services. The system may be particularly useful in rural areas where a crash is less likely to be encountered and reported quickly. Amiri et al. (2001) describe a pilot study for the Minnesota DOT's Mayday Plus notification system. The system relays the location and severity of a crash to emergency responders in the appropriate jurisdiction. Additionally, South Dakota uses GPS/GIS to enhance data elements for their emergency 911 system that includes locating structures, etc. (US Federal Highway Administration, 2000). *GPS World* (2002) reported plans by the General Motors Corporation to add automatic crash notification (CAN) services to vehicles equipped with OnStar. Currently, the OnStar system automatically notifies a commercial call center when the air bag is deployed. The location of the vehicle is determined using GPS, and relayed to a call center. The new system will send a call when the vehicle is involved in a moderate to severe crash whether or not the air bag deploys. An accelerometer can be used to determine rapid deceleration indicating a potential crash. Crash severity information will also be relayed so that the appropriate emergency services or equipment can be dispatched. General Motors Corporation plans to offer the system beginning in model year 2004.

## 6. Locating environmentally sensitive features

The location of environmentally sensitive features is necessary for location or relocation of highways. Environmentally sensitive areas may include wetland boundaries, location of threatened or endangered species, archeological or historic sites, hazard waste sites, etc. The actual physical boundaries however, may be difficult to locate. A number of environmentally sensitive features, such as endangered species habitats, are characterized by non-distinct boundaries that are not readily delineated by political or physical boundaries or identified on aerial images. It is not uncommon for agencies to maintain this type of information in formats such as hand-drawn polygons on paper maps. GPS is useful because boundaries, such as wetlands, are more easily identified by actually going into the field. GPS points can be located around the perimeter of environmentally sensitive areas, and polygons then created. Britch and Fitch (1997) evaluated the feasibility of using mapping-grade GPS to delineate wetlands in the USA for compliance with the National Environmental Policy Act (NEPA). The US Federal Highway Administration (2000) indicated that 41% of the states who responded in its study either use or are planning to use GPS to delineate wetlands and archeological site boundaries or locate sensitive features such as endangered species habitats. Other uses of GPS for environmental applications reported include:

- location of noise receptors;
- location of hazardous waste sites;
- location of illegal dumps;
- tracking of hazardous waste vehicles;
- delineation of endangered species habitat.

## 7. Summary

The widespread availability of GPS technology in recent years has fundamentally altered the way in which spatial data are collected. GPS can fairly accurately locate and report planar coordinates for a particular point as well as altitude, date, time, speed, and heading. Transportation-related GPS applications are numerous, and range in complexity from simple uses, such as collection of bus stop locations, to sophisticated uses, such as real-time vehicle tracking for ITS.

GPS is used by numerous transportation agencies including MPOs, DOTs, county and local government agencies, and transportation engineering consultants. GPS is increasingly being used for transportation applications by the private sector as well, with the advent of in-vehicle navigation systems and fleet-tracking systems. Several applications of GPS in the transportation sector have been

discussed in this chapter, and an overview of centerline mapping, inventory management, AVL, safety applications, and locating environmentally sensitive features provided.

## References

- Advantec (2002) *Hong Kong Fire Services Department Automatic Vehicle Location System (AVLS)*. Hong Kong: Advantec (<http://www.advantec-usa.com/hkavl.html>).
- Amiri, F., D. Funke and R. McClellan (2001) "Speed to the scene: GPS and sensors accelerate accident response," *GPS World*, 12:6.
- Britch, S.C. and G.M. Fitch (1997) "Opportunities for collecting highway inventory data with the Global Position System," *Transportation Research Record*, 1593:64–71.
- Burtsch, R. (2003) *Lesson 4: mobile mapping systems. Class notes*. Big Rapids: Ferris State University ([http://www.ferris.edu/faculty/burtschr/sure382/lessons\\_pdf/Lesson\\_4.pdf](http://www.ferris.edu/faculty/burtschr/sure382/lessons_pdf/Lesson_4.pdf)).
- Czerniak, R.J. and J.P. Reilly (1998) *NCHRP Synthesis of Highway Practice 258: application of GPS for surveying and other positioning needs in departments of transportation*. Washington, DC: Transportation Research Board, National Research Council.
- Drane, C. and C. Rizos (1998) *Positioning systems in intelligent transportation systems*. Boston: Artech.
- Ellis, C.D., C. Quiroga and S.-Y. Shin (2001) "Integrated platform for managing utilities along highway corridors," *Transportation Research Record*, 1768:233–241.
- GPS World (2002) *Global view – September 2002*. Eugene: GPS World (<http://www.gpsworld.com/gpsworld/article/articleDetail.jsp?id=34311>).
- Graettinger, A.J., J. McFadden and T.W. Rushing (2001) "Evaluation of inexpensive Global Positioning System units to improve crash location data," *Transportation Research Record*, 1746:94–101.
- Karimi, H.A., A.J. Khattak and J.E. Hummer (2000) "Evaluation of mobile mapping systems for roadway data collection," *Journal of Computing in Civil Engineering*, July:168–173.
- Kut, S. and D. Baraniak (2000) "Rhode Island DOT uses mobile mapping technology for asset inventory and much more," in: *Proceedings of the Geographic Information Systems for Transportation Symposium*, Presentation. Minneapolis.
- Lambda Tech (2003) *What is GPSVision™?* Waukesha: Lambda Tech (<http://www.lambdatech.com/gpsvision.html>).
- Masters, E.G., C. Rizos and B. Hirsch (1994) "GPS: more than a real world digitizer," in: *Proceedings of the 1994 IEEE Position Location and Navigation Symposium*. Las Vegas.
- McLellan, J. F., E. J. Krakiwsky and J.B. Schleppe (1993) "Application of GPS positioning to management of mobile operation," *Journal of Surveying Engineering*, 119:71–83.
- Miller, J.S. and D. Karr (1998) "Experimental application of Global Positioning System to locate motor vehicle crashes: impact on time and accuracy," *Transportation Research Record*, 1625:41–49.
- Noronha, V., M. Goodchild, R. Church, S. Kulkarni and S. Aydin (2000) *The LRMS Linear Referencing Profile: technical evaluation*. Washington, DC: Federal Highway Administration, US Department of Transportation.
- Poling, A., J. Lee, P. Gregerson and P. Handly (1994) "Comparison of two sign inventory data collection techniques for geographic information systems," *Transportation Research Record*, 1429:36–39.
- Public Works (2001) "Technology steps up to meet changing winter maintenance needs," *Public Works*, 132:20–22.
- US Federal Highway Administration (2000) *An Investigation of the use of Global Positioning System (GPS) technology and its augmentations within state and local transportation departments*, FHWA-RD-00-093. McLean: US Department of Transportation, FHWA.
- US Federal Highway Administration (2002) *Rural ITS toolbox*. Washington, DC: FHWA ([http://www.itsdocs.fhwa.dot.gov/rural\\_its.htm](http://www.itsdocs.fhwa.dot.gov/rural_its.htm)).
- Woods, R. (2002) *RoadStar GPS service rolls out*. Darien: Jupitermedia ([http://www.instantmessagingplanet.com/wireless/article.php/10766\\_1465291](http://www.instantmessagingplanet.com/wireless/article.php/10766_1465291)).

***Part 7***

**SPATIAL COGNITION**

## COGNITIVE MAPS AND URBAN TRAVEL

REGINALD G. GOLLEDGE

*University of California, Santa Barbara, CA*

TOMMY GÄRLING

*Göteborg University*

### 1. Introduction

The focus of this chapter is an examination of the relationship between cognitive maps and travel behavior in urban environments. We do this examination incrementally, beginning with clarifications of terms relating to cognitive maps, cognitive mapping, and wayfinding. We then emphasize transportation-related issues such as cognizing of transportation networks, path selection, wayfinding and navigation. We further examine problems of selecting paths to destinations by using existing transport networks. We also introduce concerns relating to the role of trip purpose in path selection, and discuss how different purposes spawn different path or route selection strategies. In a final section we briefly examine the interaction between cognitive maps, cognitive mapping, and current practice of travel choice modeling.

### 2. Basic concepts

#### 2.1. Cognitive maps

The bulk of human travel is repetitive and relatively invariant in time and space. It would be unusual for humans to consult a cartographic map of an environment prior to every trip. Rather, humans travel by virtue of the knowledge stored in their long-term memory or cognitive map. The term “cognitive map” is generally used as a metaphor or as a hypothetical construct (Kitchin, 1994). It is convenient for us to think metaphorically of consulting a “map” as we engage in developing a travel plan (Gärling et al., 1984; Gärling and Golledge, 1989). The map concept, however, is a convenient one for summarizing the processes involved in making a

geographically structured travel plan. This includes establishing locations, understanding distances between locations, comprehending the direction of one location from another, linking locations in sequence, and transferring knowledge from the mental arena to the surrounding physical environment (i.e. matching knowledge structures with street and highway networks and associated land uses).

Cognitive maps thus are the conceptual manifestations of place-based experience and reasoning that allow one to determine where one is at any moment and what place-related objects occur in that vicinity or in the surrounding space. As such, the cognitive map provides knowledge that allows us to solve problems of how to get from one place to another, or how to communicate knowledge about places to others without the need for supplementary guidance such as might be provided by sketches or cartographic maps. Traditionally, cognitive map information has been collected by asking people to produce “spatial products” or external representations of what they know about a specific place. The representations may be in the form of sketch maps, written or verbal descriptions of routes or layouts, images of places such as slides, photos, or videos, and judgments about spatial relations that might reveal a latent structural knowledge of a setting (Golledge, 2002).

## 2.2. *Cognitive mapping*

Cognitive mapping is the process of encoding, storing, and manipulating experienced and sensed information that can be spatially referenced. What guides this mental processing is being actively researched in cognitive psychology, neuropsychology, and related fields (Golledge, 1999). But, essentially, cognitive mapping involves sensing, encoding, and storing experienced information in the mind. This is referred to as “declarative” knowledge. This information is subjected to internal manipulations such as spatial thinking and reasoning. These activities manipulate and interpret the declarative knowledge base stored in long-term memory as parts of it are needed to solve problems. These problems include decision-making and choice related to travel behavior.

## 3. **Transportation issues**

### 3.1. *Cognizing transportation networks*

Most household or personal trips take place on existing transportation systems. These systems include public and private, mass and individual, modes that either share networks (e.g. cars and buses), have dedicated networks (e.g. tracked vehicles such as trains and trolley cars), or are usually confined to specific

corridors or lanes (e.g. air and surface ocean traffic). We deal here only with private and individual movements (e.g. by car, bicycle, or walking), with most emphasis on private vehicle movement.

In many countries the household car represents an important form of movement. To satisfy economy of movement, minimize air and noise pollution, achieve door-to-door delivery of drivers and passengers, and guarantee independence in route choice, networks of surface roads have been developed. Usually these are differentiated into freeways, highways, arterials (major and minor), local streets, and lanes or alleys.

When making a trip, each individual must consider how to use the local road hierarchy. These decisions can be made *a priori* (as in a travel plan) or *en route* (as in real-time wayfinding). The mere existence of the hierarchy, combined with individual memories of travel experience, leaves the way open for different route-selection strategies to be developed and for different paths to be followed. Thus one next-door neighbor might try to maximize use of a freeway for, say, a trip to work and maximize use of local streets to facilitate a trip chain on the way home, while another neighbor might use the reverse strategy. Thus, two spatially adjacent householders, going to the same destination, can choose completely different paths. By doing this, their environmental experiences may differ, and their cognitive maps thus may likewise be quite different.

In many urban environments, traffic control measures such as one-way streets and limited on-street parking can also influence path selection and, consequently, the nature of the detail that is geo-referenced in the cognitive map. Apparently, to facilitate communication and development of a general understanding of complex environments, people tend to define “common anchors” – significant places in the environment that are commonly recognized and used as key components of cognitive maps – and idiosyncratic or “personalized anchors” that are related to a person’s activities (e.g. a specific workplace or home base) (Couclelis et al., 1987). These anchor the layout or structural understanding of an environment (regardless of its scale). Objects and features in an environment “compete” for a traveler’s attention, with the most successful reaching the status of a common anchor, recognized by most people and consequently incorporated into most cognitive maps. Other features and objects are less successful in general, but might achieve salience for a specific trip purpose (e.g. “the odd-shaped building where I park in order to go to my favorite restaurant”). Minor pieces of information are attached to anchors, and act as “primers and fillers” – the second, third, or lower orders of information experienced but used only in selected ways and with varying frequencies.

Little research has been completed on the creation of network knowledge and the relationship between network knowledge systems and real-world transportation systems. We all realize from personal experience that our knowledge of existing networks is but partial and quite minimal. But, if we have an overall

anchoring structure or general layout understanding of *en route* and off-route landmarks and can determine a route or course through multiple networks of links and nodes, we can – either by using a travel aid such as a map or by independently accessing cognitively stored information – find our way between specific origins and destinations in urban environments. Sometimes this task seems simple, with minimal feasible alternative path structures to be considered (e.g. a trip from home to a nearby elementary school). At other times the task seems complex and substantial and requires serious planning and implementation (e.g. a trip from home to a distant work environment). We shall explore these concepts further below.

### *3.2. Travel behavior*

Travel behavior consists of a movement through space using a particular mode of travel. It can be recorded as a trace throughout the environment. This trace is sometimes called a path or a route. Paths or routes are defined by selecting sections from a network of connecting nodes and links. Nodes consist of places where links join or intersect. The route then consists of a sequence of links and nodes between a specified origin and destination. Different human activities require designation of different routes in order to link places where wants or needs can be satisfied. Routes must be experienced and learned if they are to be used repeatedly over time. Learning a route involves identifying the origin and destinations, knowing the number of link segments and their appropriate sequencing; recognizing intersection nodes and identifying choice points where turning decisions may have to be made; remembering the number and direction of turns embedded in a given route; being able to recognize on- or off-route landmarks that help interpret where one is along the route at any particular point in space or time; and being able to retrace and/or reverse the route on an as-needed basis. If the route is circuitous, the learning process will involve understanding its configurational complexity, thus facilitating the process of taking a shortcut if needed (e.g. if the route is blocked by congestion, construction, accident, or some other barrier).

Sholl (1996) suggests that travel requires humans to activate two processes that facilitate spatial knowledge acquisition – person-to-object relations that dynamically alter as movement takes place, and object-to-object relations that remain stable even when a person undertakes movement. The first of these is called egocentric referencing; the second is called the anchoring structure of a cognitive map (or layout referencing). Given this conceptual structure, it is obvious that poor person-to-object comprehension can explain why a traveler can become locally disoriented even while still comprehending in general the basic structure of the larger environment through which movement is taking place. Error in encoding local and more general object-to-object relations can result in

misspecification of the anchor point geometry on which cognitive maps are based. The latter seems to be responsible for many of the distortions and fragmentations found in attempts to externalize cognitive maps (frequently referred to as "spatial products," see Liben, 1981).

In both cognitive mapping and wayfinding, environmental anchors play an important role. Anchors can be landmarks (on- or off-route), important choice points (e.g. transfer between a freeway and arterial or local streets), path segments (e.g. the final freeway segment before exiting to work or home), or even a distinct area (such as a park, shopping center, or ethnic or cultural neighborhood). Their actual cognized locations and the awareness of the spatial relations among them (i.e. their layout) provide a framework on which is grafted piecemeal knowledge acquired during urban experience (e.g. personal travel, television or video or film coverage, or verbally described places in the city).

Although there are many electronic, hard copy, and other technical aids that can be used as wayfinding tools, nevertheless, humans most frequently tend to use cognitive maps and recalled information as travel guides. There are three different types of knowledge usually specified with relation to travel behavior. Perhaps the most common is called route learning (or systematic encoding of the route geometry by itself). A second concept is route-based procedural knowledge acquisition that involves understanding the place of the route in a larger frame of reference, thus going beyond the mere identification of sequenced path segments and turn angles. A third version is called survey or configural knowledge, and this implies comprehension of a more general network that exists within an environment and from which a procedure for following a route can be constructed.

Human-based methods for wayfinding carry all the imprecision and error baggage that instruments were designed to eliminate. This error baggage includes a variety of spatially based concepts. For example, many studies show that human pointing errors (even between familiar places) average about 25–30°. In addition, shorter distances are usually overestimated while longer distances are underestimated. Perceived distances to and from a particular origin and destination are often perceived to be asymmetric. Triangle inequality does not always hold for judged distances between places. People do not always perceive the same object to be at the same place. And changing perspective often changes the evaluation of spatial relations (e.g. with regard to left/right, front/back, up/down, and along/across). It can be expected, therefore, that spatial representations in humans are incomplete and error-prone, producing the distortions or fragmentations of spatial products that have been found by numerous researchers. But what is significant of course is that an individual need not have a correctly encoded and cartographically correct "map" stored in memory to be able to successfully follow a route. Route knowledge by itself requires that a very small section of general environmental information is encoded. In its pure form, the route is completely self-contained, anchored by choice points and *en route* landmarks and consisting

of consecutive links with memorized choice points and turn angles between the links. The integration of specific routes is a difficult task, but apparently not an impossible one, for many people develop either skeletal or more complete representations of parts of urban networks through which their episodic travel takes place (Ishikawa, 2002).

### 3.3. Path selection criteria

Human wayfinding can be regarded as a purposive, directed, and motivated activity that may be observed and recorded as a trace through an environment. The trace is usually called the route or course. A route results from implementing a travel plan (Gärling et al., 1984; Gärling and Golledge, 1989) which is an *a priori* decision-making process that defines the sequence of segments and turn angles that comprise the course to be followed or the general sector or corridor within which movement should be concentrated. The route represents the trace over the ground (spatial manifestation) from following a specified course. The travel plan is the outcome of using a particular strategy for path selection.

A large number of different criteria are used in path selection. The major types that can be found referred to in travel-related literature in fields such as travel behavior, operations research, transport geography, and behavioral travel modeling are summarized in Box 1.

### 3.4. Navigation and wayfinding

It is becoming more common to differentiate between navigation and wayfinding. Navigation implies that a route to be followed is predetermined, is deliberately calculated (e.g. humans often use mechanical equipment and mathematical equations to do this), and defines a course to be strictly followed between a specified origin and destination. Progress along the course is sometimes monitored (e.g. by air traffic controllers or, in the case of private delivery systems such as UPS or FedEx, by centralized tracking of vehicles using the Global Positioning System (GPS)). Wayfinding is taken more generally to involve the process of finding a path (not necessarily previously traveled) in an actual environment between an origin and a destination that has previously not necessarily been visited. Wayfinding can thus be identified with concepts such as search, exploration, and with incremental path segment selection during travel. Wayfinders can also use technical assistance (e.g. a compass, GPS, or a network map) but, more often, use cognitive maps.

Navigation is usually dominated by criteria such as shortest time, shortest path, minimum cost, least effort, or with reference to specific goals that should

Box 1  
Types of route selection criteria

- Longest leg first
- Shortest leg first
- Fewest turns
- Fewest lights or stop signs
- Fewest obstacles or obstructions
- Variety seeking behavior
- Minimizing negative externalities (e.g. pollution)
- Avoiding congestion
- Avoiding detours
- Responding to actual or perceived congestion
- Minimizing the number of segments in a chosen route
- Minimizing the number of left turns
- Minimizing the number of non-orthogonal intersections
- Minimizing the number of curved segments
- Ensuring locomotion remains within a given width (corridor) surrounding a straight line connection between origin and destination
- Maximizing aesthetics
- Minimizing effort
- Minimizing actual or perceived cost
- Minimizing the number of inter-modal transfers
- Minimizing the number of layers of a road, street, or highway system that have to be utilized
- Fastest route
- Least hazardous in terms of known accidents
- Least likely to be patrolled by authorities
- Minimizing exposure to truck or other heavy freight traffic

be achieved during travel. Thus, it lends itself to optimization modeling. Wayfinding is not as rigidly constrained, is purpose-dependent, and can introduce emotional, value, and belief considerations, and satisfying constraints into the travel process. This procedure lends itself to stochastic probability models or any of a variety of logistic models. Whereas navigation usually requires the traveler to preplan a specific route to be followed, wayfinding can be more adventuresome and exploratory, without the necessity of a pre-planned course that must be followed. While for some purposes travel behavior will be habitualized (thus lending itself to the optimization modeling activities of the navigation process), for other purposes variety in path selection may be more common (indicating more of a wayfinding concern and requiring a different type of model base).

### 3.5. Route learning

Repeated path following facilitates remembering path components and recalling them for further use. This is called route learning. Paths or routes are represented

as one-dimensional linked segments or, after integration with other paths, as networked configurations. The latter, along with on- and off-route landmarks, spatial relations among them, and other spatial and non-spatial attributes of places – such as prominence of visible form – make up the anchoring layout of a remembered environment. Route-learning and route-following strategies help build up cognitive maps via an integration process (Gärling et al., 1981; Ishikawa, 2002). Difficulties experienced in mentally integrating different routes and their associated features into network structures help to explain why cognitive maps may be fragmented, distorted, and irregular (Gale et al., 1990).

### *3.6. The role of trip purpose*

Human wayfinding is very trip-purpose-dependent, and it is thus difficult to attribute any specific cognitive process to wayfinding generally. The question remains as to whether specific purposes are better served by certain types of wayfinding strategies. For example, the journey to work, the journey to school, and the journey for convenience shopping are often best served by quickly forming travel habits over well-specified routes. Such an activity would minimize *en route* decision-making, and often it conforms to shortest path principles. However, journeys for recreation or leisure may be undertaken as search and exploration processes, and require constant locational updating and destination fixing. Thus, as the purpose behind the activity changes, the path selection criteria can change, and, as a result, the path that is followed (i.e. the travel behavior) may also change. Recent work on intelligent highway systems (IHS) and advanced traveler information systems (ATIS) has shown that humans sometimes respond to advanced information on congestion or the presence of obstacles by substituting destinations, by changing travel times (particularly in the early morning), by delaying or postponing activities, or by selecting alternate routes (particularly in the evenings) (Chen and Mahmassani, 1993). All these produce different travel behaviors in response to trip purpose changes. Cognitive maps must be very versatile to allow such behavioral dynamics.

### *3.7. Travel guidance*

To help minimize inefficiencies in travel behavior that contribute to excess air pollution, noise, and danger, IHS are being developed to provide advanced and *en route* information for the upper levels of road hierarchies (e.g. freeways, highways, and major arterials). These include:

- pre-trip information on traffic conditions (speed, congestion, and accidents);
- variable message signs and other automated traffic management systems;
- *en route* radio broadcasts;
- automated vehicle guidance systems (e.g. GPS or local positioning system locators and in-car route maps);
- automated transit advisory systems;
- directional lane control;
- specialized traffic lanes (e.g. bus lanes, carpool lanes).

#### 4. Incorporating cognitive maps into travel choice models

Models that link spatial behavior to travel choice have over the past 30 years become more dominant (McFadden, 2001). Ways to incorporate spatial cognition in such models have not yet, however, achieved widespread adoption by transportation professionals interested in predicting flows over a transportation network, presumably because of their difficulties in operationalizing and measuring constructs such as cognitive maps, cognitive mapping, navigation, and wayfinding. Yet, no one denies that travelers' choices depend on what spatial information they perceive and store in long-term memory.

Current opinion appears to indicate that, because factors such as cognitive-mapping ability, cognitive map knowledge of feasible alternatives, navigation and wayfinding strategies, and preferences for path selection criteria all are presumed to have a substantial impact on travel choices, there is a growing need to include spatial cognition explicitly in models. Specifically, cognitive maps must become a part of the modeling process in that they are summaries of what is known about the network over which travel must take place; they provide information on what is known about the location, possible destinations, and feasible alternatives for any choice; and they provide a means for spatializing attribute information by attaching values and belief or preference ratings or measures to specific geocoded places.

In an attempt at addressing these issues, Gärling and Golledge (2000) posit that information stored in the cognitive map impacts travel choices in that (1) potential travelers can only choose from known destinations brought to their attention, and (2) knowledge of spatial relations between these destinations impact choices of them as well as choices of travel between them. Furthermore, the degree and accuracy of knowledge dependent on familiarity with the environment are important moderators of these relationships. In particular, it is argued that multistop multipurpose trip making requires an extensive and accurate cognitive map.

Promising attempts to incorporate cognitive maps and cognitive mapping in travel choice models are perhaps most evident in computational process models

(Smith et al., 1982; Gärling et al., 1994). In particular, Gärling et al. (1989, 1998) make explicit assumptions about the role of these factors. However, empirical verification is still needed. A step in the direction of producing comprehensive models that incorporate issues of cognitive concepts is that taken by Arentze and Timmermans (2002), who are developing a rule-based model capable of learning the environment. We believe that important progress can be expected in these respects.

Two early computation process models stand out for their attempts to incorporate cognitive map concepts. These are Scheduler (Gärling et al., 1989) and GISICAS (Kwan, 1994). Scheduler required a potential traveler to indicate places and time slots that would be filled with obligatory activities. Discretionary events (e.g. going to the gym, shopping or socializing) were then integrated into a person's activity schedule, depending on the time required to perform an activity, the travel time to and from the activity place, and a subjective priority attached to the activity. GISICAS built on Scheduler by using geographic information system procedures to define a "feasible set" of alternative locations where activities could be performed, producing a map representation of the alternatives with most probable routes plotted. The feasible alternative procedure was an operationalization of the cognitive map idea. It defined a small set of alternative locations that were probably "best known" to a given traveler. This concept was based on proximity to home, work locations, and to places within a pre-set distance corridor along the most likely route that linked home and work.

An even more difficult task is to provide a cognitive map measure for incorporation into logistic choice models – the most favored format for analyzing and predicting travel associated with daily activity patterns. Ben-Akiva et al. (2001) are experimenting with a set of scale values representing a person's self-assessed spatial ability. Factors such as self-assessed ability to estimate direction and distance, self-assessed knowledge of landmark layout in an area, and self-assessed ability to perform spatial tasks such as wayfinding, recalling distant places, and so on are being evaluated to see how well people evaluate local spatial knowledge structures. Eventually, the results may be incorporated as latent structure variables in travel behavior models.

## **5. Conclusion**

Because of individual differences in the content of cognitive maps, different motivations or purposes for travel, and different preferences for optimizing or satisfying decision strategies, human travel behavior is difficult to understand or predict. If we add to that the unexpected barriers and obstacles to traffic flow that occur spontaneously and intermittently (e.g. from congestion, accidents, construction, or other obstacles that impede movement over a selected path or

over a network), then problems of intelligently modeling travel behavior in the real world become substantial. Yet, some success has been achieved in doing this, using simplified assumptions about human behavior (e.g. assuming that, knowingly or unknowingly, travelers adopt shortest-path optimizing practices). But models like this and the predictions they make can be very inadequate. The question facing future research is that of combining travel demand (considering people's activities) with network supply (considering the tracks, corridors, or transport systems available) with an understanding of how humans decide on where they prefer (or have) to go and how they prefer (or have) to get there. Emphasizing cognitive-mapping principles may give a level of insight that has not so far been provided.

## References

- Arentze, T. and H.P.J. Timmermans (2002) "Modeling learning and adaptation processes in activity-travel choice: a framework and numerical experiments," *Transportation*, 30:37-62.
- Ben-Akiva, M., S. Rammig and R.G. Golledge (2001) *Collaborative research: individuals' spatial behavior in transportation networks*. National Science Foundation Grant Proposal BCS-0083110. Santa Barbara: University of California and Massachusetts Institute of Technology.
- Chen, P.S.T. and H.S. Mahmassani (1993) "A dynamic interactive simulator for studying commuter behavior under real-time traffic information supply strategies," *Transportation Research Record*, 1413:12-21.
- Couclelis, H., R.G. Golledge, N. Gale and W. Tobler (1987) "Exploring the anchor-point hypothesis of spatial cognition," *Journal of Environmental Psychology*, 7:99-122.
- Gale, N.D., R.G. Golledge, J. Pellegrino and S. Doherty (1990) "The acquisition and integration of neighborhood route knowledge in an unfamiliar neighborhood," *Journal of Environmental Psychology*, 10:3-25.
- Gärling, T. and R.G. Golledge (1989) "Environmental perception and cognition," in: E.H. Zube and G.T. Moore, eds, *Advances in environment, behavior, and design*, Vol. 2. New York: Plenum Press.
- Gärling, T. and R.G. Golledge (2000) "Cognitive mapping and spatial decision-making," in: R. Kitchin and S. Freundschuh, eds, *Cognitive mapping: past, present and future*. London: Routledge.
- Gärling, T., A. Böök, E. Lindberg and T. Nilsson (1981) "Memory for the spatial layout of the everyday physical environment: factors affecting the rate of acquisition," *Journal of Environmental Psychology*, 1:263-277.
- Gärling, T., A. Böök and E. Lindberg (1984) "Cognitive mapping of large-scale environments: the interrelationship of action plans, acquisition and orientation," *Environment and Behavior*, 16:3-34.
- Gärling, T., K. Brännäs, J. Garvill, R.G. Golledge, S. Gopal, E. Holm and E. Lindberg (1989) "Household activity scheduling," in: *Transport policy management and technology towards 2001: selected proceedings of the Fifth World Conference on Transport Research*, Vol. IV. Ventura: Western Periodicals.
- Gärling, T., T. Kalen, J. Romanus, M. Selart and B. Vilhelmsen (1998) "Computer simulation of household activity scheduling," *Environment and Planning A*, 30:665-679.
- Gärling, T., M.P. Kwan and R.G. Golledge (1994) "Computational process modelling of household activity scheduling," *Transportation Research B*, 28:355-364.
- Golledge, R.G., ed. (1999) *Wayfinding behavior: cognitive mapping and other spatial processes*. Baltimore: Johns Hopkins University Press.
- Golledge, R.G. (2002) "Cognitive maps," in: K. Kempf-Leonard, ed., *Encyclopedia of social measurement*. San Diego: Academic Press.
- Ishikawa, T. (2002) "Spatial knowledge acquisition in the environment: the integration of separately learned places and development of metric knowledge," unpublished Ph.D. thesis. Santa Barbara: University of California.

- Kitchin, R.M. (1994) "Cognitive maps: what are they and why study them?" *Journal of Environmental Psychology*, 14:1-19.
- Kwan, M.P. (1994) "A GIS-based model for activity scheduling in intelligent vehicle highway systems (IVHS)," unpublished Ph.D. thesis. Santa Barbara: University of California.
- Liben, L.S. (1981) "Spatial representation and behavior: multiple perspectives," in: L.S. Liben, A.H. Patterson and N. Newcombe, eds, *Spatial representation and behavior across the lifespan*. New York: Academic Press.
- McFadden, D. (2002) "Disaggregate behavioral travel demand's RUM side: a 30-year retrospective," in: D.A. Hensher and J. King, eds, *The leading edge in travel behavior research*. Oxford: Pergamon Press.
- Sholl, M.J. (1996) "From visual information to cognitive maps," in: J. Portugali, ed., *The construction of cognitive maps*. Dordrecht: Kluwer.
- Smith, T.R., J.W. Pellegrino and R.G. Golledge (1982) "Computational process modelling of spatial cognition and behavior," *Geographical Analysis*, 14:305-325.

*Chapter 29*

## SPATIAL PROCESSES

RYUICHI KITAMURA

*Kyoto University*

### 1. Introduction

Spatial processes – behavioral processes that evolve in space – are at the heart of travel demand analysis and transportation planning. Spatial processes evolve at different spatial scales: consider inter-regional migration, interchange of telephone calls, commuting in a metropolitan area, trip distribution for grocery shopping, or movements of visitors in a shopping mall. Among these, the focus of this chapter is on residents' daily movements in an urban area. Following the nomenclature of the field of travel behavior research, the term, "travel," is hereafter used in place of "movement," and movement from an origin to a destination, where an activity is typically engaged, will be called a "trip." This chapter is thus concerned with the daily travel patterns of urban residents.

The spatial process evolves over time. It is "stochastic" in the sense that it comprises random events that materialize over time. In the case of daily travel patterns, an event may be defined as the engagement in an activity at a different location, which induces a trip. Thus, a person's daily travel pattern typically contains several events and trips that connect them. In the sense that the timing, destination, and other attributes of the trip, and the type and duration of the activity, are not known to the observer beforehand, they may be viewed as random events.

Describing travel patterns in space may not appear overly complicated: once one knows how the origins and destinations of travel are distributed spatially, and how the origins and destinations are connected by transportation networks, then it should be a relatively easy task to find the resultant spatial pattern of travel. In fact, many principles have been proposed for this application, and have produced reasonable results. This apparent simplicity of analyzing spatial patterns, however, is due to the assumption that each trip can be isolated and analyzed individually, disregarding movements that are connected to it. Whereas an individual's movement is continuous in space, there is always a trip prior to any trip, and sooner or later there will be another trip after it; and the destination of a

trip is the origin of the next trip. Obviously, the series of trips are not independent of each other.

The limitations of trip-based analyses have been well articulated in the literature, particularly in the field of activity-based analysis of travel behavior (e.g. see Jones et al., 1990; Ettema and Timmermans, 1997). Past studies on trip chaining<sup>a</sup> have shown how the purposes of trips tend to be linked and sequenced in a trip chain. It has also been shown that the destination of a non-home-based trip<sup>b</sup> is influenced by the location of the home base; therefore, the spatial distribution of trips cannot be determined by the attributes of their origins and destinations and the separation between them. Furthermore, defining the attraction of a destination is not a straightforward matter when one considers the possibility that the traveler may chain trips to visit multiple locations in geographical proximity. Empirical evidence also supports the theoretical notion that the duration of a trip is correlated with the time spent at the destination, and indicates that the sensitivity to spatial separation exhibited in the choice of destination varies by the time of day. A more holistic framework is desired to analyze the series of trips made by a traveler as a whole, rather than each trip individually outside of the context in which it was made.

Such frameworks have been proposed by Chapin, Hägerstrand, and their co-workers. Chapin (1978) notes that to "understand how patterned forms of human activity evolve, it is essential to begin with the behaviour of individuals at the micro-level of daily routines," and focuses on motivations for activity engagement on the ground that "activity choice itself can only be explained in terms of the motivation, needs, wants and capabilities of the individual" (Hemmens, 1970). The decision to engage in an activity implies the allocation of time and the selection of location for that activity. Chapin and his colleagues have attempted to represent human activity with respect to time, space, and activity, given opportunities that are distributed over space. Hägerstrand's focus, on the other hand, is on constraints, and therefore possibilities, that govern individuals' activities: "A major emphasis in the time geographic approach has been placed on analysing the possibilities open to the individual and not merely his actually chosen and observed behaviour" (Lenntorp, 1978). These two streams of research underlie the activity-based analysis of travel behavior mentioned earlier.

Yet, analyses of spatial aspects of travel are rather rare and limited in the field of travel behavior analysis. This is at least in part due to the difficulties associated with the representation of trips and opportunities in space. In this chapter, selected studies of spatial processes are reviewed, with emphasis on how spatial

<sup>a</sup>The trip chain refers to the series of trips that originates and terminates at a base (e.g. the home, workplace, or school). The term "tour" is used interchangeably with trip chain.

<sup>b</sup>A trip is described as "non-home-based" when neither its origin nor destination is the home base.

elements are represented in these studies. We begin this with trip-based studies, then move on to studies of trip-chaining behavior. Following these, simulation models of spatial processes are reviewed.

## 2. Trip-based studies and their limitations

Daily travel patterns as a special case of spatial processes are complex and difficult to represent (Burnett and Hanson, 1979). The complexity stems in part from the fact that multiple trips are linked to form a travel pattern. The decisions associated with the respective trips are thus interdependent. Travel patterns also reflect the traveler's act of planning and his or her reactions to unexpected events; they are subject to the cognitive limitations of the traveler, and are influenced by a variety of constraints, many of which are often unobserved in the analysis of travel patterns. The complexity is also due to the fact that there are numerous – theoretically infinite – alternative daily patterns that an urban resident may adopt; and the fact that some travel patterns are highly random, at least to the eye of the analyst, while others are quite regular.

Also difficult is the representation of space in the analysis of travel patterns. The spatial expanse of an urban area is typically represented by a set of geographical zones, and the spatial attributes of a trip are defined in terms of the zones where its origin and destination belong. This, however, may not offer the required level of spatial resolution when the number of zones is small. When the number of zones is large, on the other hand, it becomes difficult to comprehend spatial patterns expressed as an origin-to-destination (O-D) trip matrix or as desire lines that link origins and destinations on a zone map. It is then quite understandable that a summary representation of spatial aspects is often deployed in the analysis of travel patterns, such as the distribution of trip lengths, or the distribution of broadly classified destination locations – e.g. the central business district (CBD) versus others.

Representing the spatial distribution of opportunities is another issue. This is often done using some measure of accessibility as an indicator. Yet, it is not immediately clear how accessibility is related to travel patterns that involve linked multiple trips. For example, “very little is known about the impact of relative accessibility on trip frequencies, trip-purpose mixes, or destination choices at the aggregate level, and ‘even less is known about [the] implications [of accessibility] at the level of individual or household behavior’” (Wachs and Koenig, 1979). Furthermore, the relationship between accessibility and spatial patterns does not appear to have been well studied. Only recently has some basic analysis been performed on how spatial patterns are influenced by spatial structure (Timmermans et al., 2003).

The analysis of spatial processes can be immensely simplified once one accepts the assumption that a trip can be analyzed as an independent entity. This is the basis of the conventional trip distribution analysis. More specifically, "Movement is normally related to three basic variables: a measure for distance between the origin and the destination; a measure for the attraction of the destination, and a measure for the movement potential at the origin, e.g. socio-economic characteristics of households" (Kofoed, 1970). This can be typically seen in the gravity model of trip distribution.

The literature is rich with models of spatial interaction, e.g. the gravity model, entropy model, intervening opportunities model, utility model, minimal information model, and discrete-choice model. Despite the variety of disciplines from which these models draw their basic relations – Newtonian physics, probability theory, statistical physics, information theory, and micro-economics – equivalence among them has been shown in a number of studies (e.g. Wilson, 1970; Beckmann and Golob, 1972). In particular, the multinomial logit model of destination choice (e.g. Lerman, 1976), which is based on the principle of utility maximization, can be viewed as an application of the production-constrained gravity model (for the latter, see Wilson, 1970). In addition to offering another theoretical basis to the gravity model of trip distribution, the advent of discrete-choice analysis has facilitated the incorporation of the multitude of contributing elements – travel mode and level-of-service, as well as attraction of destination location – into measures of accessibility (Ben-Akiva and Lerman, 1979).

Nonetheless, representing the spatial distribution of trips or opportunities in space is not a trivial task. Although a set of geographical zones assigns a numbering system to the expanse of urban space, it does not represent space itself. As noted above, trip interchanges summarized as an O-D table are quite often incomprehensible when the number of zones exceeds a few dozen. In this context, the ability of spatial interaction models to offer parameters that represent tendencies in spatial patterns in a highly condensed manner emerges as a strong advantage. The most typical example is the distance decay function and its parameters of the gravity model, which summarily indicate how spatial interaction decreases with the separation between the origin and destination.

Quite often, however, the parameters of spatial interaction models, including the distance decay parameters, are found to be unstable over time or across areas (e.g. Elmi et al., 1999). Instability in parameter values may be caused by "dominant trends" affecting consumer behavior and service location, e.g. growing mobility due to improvements in the transport network and vehicles, in particular increased private car ownership, or the increased proportion of married women who go out to work (Mikkonen and Luoma, 1999).

The issue of stability cannot be fully examined by just examining observed patterns of spatial interaction. As noted earlier, gravity models of movement were "adapted from the Newtonian concept of gravitation" (Kofoed, 1970). The law of

gravity in physics is essentially static and holds for any time of observation. Travel, on the other hand, comprises discrete events that take place with certain frequencies. Kofoed continues: "when the interaction between the frequency of activity and the distance to activity locations differs from one activity to another, the comparison of the number of counted trip purposes with measured trip lengths seems to be of dubious validity. ... Hence, distance as a social factor cannot be expected to relate to human spatial activities in a simple manner and any concept should take this into account." As urban structure changes with evolving location patterns and transportation networks, the distance to activity locations changes: socio-economic and technological changes are likely to prompt changes in trip frequencies. There is no reason, then, to assume that the parameters of trip distribution models will remain stable over time.

Another reason why conventional trip distribution models may not be adequate is the fact that trips are not independent. The individual's movement pattern over a period, say a day, comprises linked trips that collectively form a continuous trajectory in space. Although a majority of individuals may exhibit "simple" travel patterns of leaving home, visiting a single location, then returning home, a large portion of trips belong to "complex" travel patterns that involve multiple destinations that are connected by trips in many possible ways (Hanson, 1979). Recognition of this led to the analysis of trip-chaining behavior, which is reviewed in the next section. In the remainder of this section, the competing destinations model proposed by Fotheringham is briefly reviewed, as it is relevant to the analysis of trip chaining.

The competing destinations model was proposed as a solution to the map pattern problem, namely the problem that the estimate of the distance decay parameter of a spatial interaction model is biased because it is influenced by the spatial distribution of opportunities. Fotheringham (1983) proposed the introduction of an accessibility measure into the model of spatial interaction, i.e.

$$T_{ij} = A_i O_i W_j^{\gamma_i} d_{ij}^{-\beta_i} H_j^{\delta_i},$$

where  $T_{ij}$  is the magnitude of spatial interaction (e.g. the number of trips) between  $i$  and  $j$ ,  $A_i$  is a scaling constant,  $O_i$  is the total number of trips originating from  $i$ ,  $W_j$  is the measure of attraction for  $j$ ,  $d_{ij}$  is the measure of separation between  $i$  and  $j$ ,  $H_j$  is an accessibility measure for  $j$ , and  $\gamma_i$ ,  $\beta_i$ , and  $\delta_i$  are parameters. The measure  $H_j$  represents the accessibility for all possible destinations that can be reached from  $j$ , and may be defined as

$$H_j = \sum_{k \neq j} P_k / d_{jk}^{\sigma},$$

where  $P_k$  is an attraction measure for  $k$ , and  $\sigma$  is a parameter. When the parameter  $\delta_i$  takes a positive value, it represents the agglomeration effect produced by the

opportunities around  $j$ , and when it takes a positive value, it represents the competition effect.

The competing destinations model was proposed for long-distance, inter-regional movements where, Fotheringham proposed, the traveler chooses a regional cluster of destinations first, then chooses a particular destination within the cluster. A similar conceptualization applies to daily travel within an urban area; for example, a shopper may compare downtown and suburban shopping centers first, then choose a particular store within the chosen cluster.

The formulation of the competing destinations model is also appealing when trip-chaining behavior is considered. If a traveler plans to visit multiple locations, then it can be reasonably assumed that his or her choice of a first destination will take into consideration the accessibility from the first destination to potential second destinations. This can be represented by incorporating the accessibility measure in a similar manner as in the competing destinations model. In the field of travel behavior analysis, a destination choice model has been proposed by Kitamura (1984) to account for the possibility of trip chaining. This model may be viewed as a version of the competing destinations model.

### 3. Trip-chaining analyses

When trips are linked, spatial interaction can no longer be examined by simply inspecting the origin and destination ends of a trip and the spatial separation between the two. For example, based on the coefficient estimates and their statistical significance in multinomial logit destination choice models for non-home-based trips, Kitamura et al. (1998) report that the travel time from a potential destination to the home base is as important in destination choice as the travel time from the trip origin to the potential destination. In addition, the series of destinations in a trip chain are all interrelated because of the simple fact that the destination of a trip is the origin of the next trip.

Trip chaining also influences, and is influenced by, the choice of travel mode, which in turn interacts with the choice of destination. There are obvious constraints on mode choice; for example, if a traveler has not left home by private car, it is very unlikely that he or she will be driving it before returning home. In fact the transition between travel modes in a trip chain is dominated by transitions from a mode to itself, leading to homogeneity in travel mode (Kitamura et al., 1997)<sup>a</sup>. It is

<sup>a</sup>The analysis is based on conventional travel survey data where movements within the same premises, such as a shopping mall, are not considered as trips. Otherwise, the transition between driving and walking would be more prevalent within a trip chain.

then obvious that the choice of mode for a trip in a trip chain cannot be examined by comparing the travel modes that compete between the origin and destination of that particular trip.

Furthermore, in the case where multiple stops are contained in a trip chain, there is no obvious definition of the distance traveled by activity (Hanson and Hanson, 1981), "travel time to an activity" (e.g. Golob, 2000), or the travel time ratio. The distance traveled by activity or travel time to an activity refers to the amount of travel made to engage in a particular type of activity. The travel time ration, proposed by Dijst and Vidaković (2000), is defined as the amount of time spent to reach an activity site divided by the amount of time spent for the activity. These concepts rest on the assumption that the amount of time spent traveling to engage in an activity can be unambiguously determined, which is not the case for trip chains with multiple stops.

For example, consider the case where a commuter visits a grocery store for a loaf of bread on the way home from work. Suppose he travels 5 min along his usual commute route and stops at a grocery store, purchases a loaf of bread, then travels another 25 min to home. Alternatively, suppose he travels 25 min to another grocery store on his commute route, purchases an identical loaf of bread, then travels another 5 min to home. And suppose it takes 30 min if the commuter drove home directly from work. Should the travel time to this shopping activity be defined as 25 min, 5 min, or 0 min? Essentially the same question arises for the definition of travel time ratio. The notion of travel time to activity or travel time ratio thus begs the question of how the amount of time spent for traveling can be divided among activities of different types when multiple types of activities are pursued in a multi-stop trip chain.

There are also issues of how activity locations are organized into trip chains, how activities are sequenced within the chain, and how all these are related to constraints associated with activity engagement and travel and also with the fixities of the respective activities. It is conceivable that there are hierarchies of activities, from mandatory, important and/or fixed, to discretionary and/or flexible. The location and timing of less-fixed activities may be subordinate to those of more fixed activities. The locations, sequence, and timing of the stops in a trip chain are determined under a myriad of constraints and a complex system of preferences. Organizing stops into trip chains and choosing suitable travel modes are additional dimensions involved in the decision underlying trip chaining. The spatial processes are thus closely associated with trip chaining, analyzing which is indeed "complex" (Burnett and Hanson, 1979).

Early studies of trip-chaining behavior were concerned with the transition of trip purposes or land uses at destinations in trip chains, i.e. exploring the characteristics of the series of trip purposes or land uses in a trip chain. Markov chains have quite often been applied to describe the linkages (e.g. Nystuen, 1967; Horton and Wagner, 1969; Hanson and Marble, 1971; Kondo, 1974).

The study by Wheeler (1972) addresses the spatial interaction of trips in trip chains, represented by a set of 60 tracts for each purpose. Wheeler notes that "multiple-purpose trips, though tied to residential locations scattered throughout the study area, most typically linked with the more centrally located tracts having employment, business, and commercial activities." The examination of "salient" flows, i.e. flows between tracts that are substantially greater than the statistical expectations, on the other hand, has found that areas with "lower population densities and relatively fewer business and commercial activities, were disproportionately involved in linking trip purposes. It was here, at greater distances from the city center, that the advantages of linking trips may be particularly great, and a higher proportion of all trips had more than a single trip" (Wheeler, 1972).

Marble and Bowlby (1968) examined the records of stops made by 116 households from Cedar Rapids, Iowa, over a 30 day period. In their study, a sample trajectory of a household over a 30 day period is shown in terms of airlines connecting the destinations visited. The spatial distributions of residences or retail opportunities are shown with a resolution of at least 0.1 mile. In the analysis, however, spatial patterns are reduced to the distribution of distances from home and measures of repetitive visits to the same opportunities. In their analysis of "variation in zonal interchange," Horton and Wagner (1969) adopted a simple zone system comprising the CBD and four quadrants. Their findings include: "Common to all three classes is a high intrazonal percentage of movement, except in the CBD area" and "The housewife-student class shows an even greater tendency to restrict movement to within zones." These are some of the few empirical studies in the literature that have addressed the spatial aspects of trip-chaining behavior.

Earlier studies of trip chaining are thus concerned primarily with the transition of trip purposes and land uses, and studies that address spatial aspects are rather scarce. The scope of trip-chaining analysis has been extended to including the sequencing of activities and history dependence (Kitamura, 1983), and patterns in which stops are organized into chains (Strathman et al., 1994). Markovian analyses have been extended to include a continuous time dimension (Lerman, 1979), or to critically examine basic assumptions, to achieve a more realistic representation of trip-chaining behavior. Notable are the examinations of the assumptions of time homogeneity (or stationarity) (O'Kelly, 1981) and history (or future) independence (Kitamura, 1983) in activity transition.

Markovian studies represent trip chaining as stochastic processes, just like a frog jumping from lily pad to lily pad. They do not attempt to capture the decision mechanism underlying trip-chaining behavior. An early study that attempted to model the decisions underlying trip chaining is by Marble (1967), who adopted the framework of game theory. O'Kelly (1981) formulated retail location as an equilibrium problem while taking trip chaining into account. Attempts to

formulate the decision to chain trips as a rational decision can be found in micro-economic studies that adopt simple conceptual frameworks of multi-stop shopping, where the cost of travel, purchase price, and inventory cost are balanced (e.g. Thill, 1985). These studies, however, rest on highly idealized settings with a variety of simplifying assumptions. Thill and Thomas (1987) note that “the theoretical approach has proved valuable to a formal conceptualization of trip chaining” but “general microeconomic frameworks do not yet allow us to deal successfully with spatial problems faced by geographers: trip-chaining structure ... or central place structures.”

In addition to these approaches there have been developments along several other lines of analysis. Random utility models have been applied to trip chaining (e.g. Adler and Ben-Akiva, 1979; Kitamura, 1984), some viewing trip chaining as sequential decisions, with others viewing it as a decision concerned with the entire pattern of trip chaining. A system of linear equations has been proposed to link trip generation and trip chaining (Goulias et al., 1990), and a structural equations model has been proposed to link time use and trip chaining (Golob, 2000). Day-to-day variations in trip chaining are addressed by Jou and Mahmassani (1997). Spatial patterns, however, are not well addressed in these studies. Indeed, the treatment of spatial elements tends to be tenuous at best in most studies.

In one of the first studies to address the relationship between spatial patterns and spatial setting, Timmermans et al. (2003) note that the “central focus of this article concerns the relationships between spatial context and the complexity of travel patterns. Complexity in this paper refers to the number of trips and tours made.” Complexity in spatial patterns is thus not of interest here. Daily travel patterns are characterized in the study in terms of “the average number of daily home-based tours per person by day of the week,” “the average number of daily trips per person by day of the week,” and “the average daily trip/tour ratio by day of the week.” Location is classified as “(i) urban areas, with good transport facilities, (ii) urban areas with bad transport facilities, (iii) suburban areas within a public transport corridor, (iv) suburban areas not within a public transport corridor and (v) the country-side.” Spatial setting is thus represented by a set of area categories defined in terms of the degree of urbanization and transportation development.

As this brief review of studies on trip-chaining behavior has suggested, studies that address the spatial aspects of trip chaining are rare. This may be due to difficulties in representing location meaningfully in the analysis of trip chaining. Those few studies that have addressed spatial elements have done so by adopting a very simple zone system, by reducing the dimensionality of the problem by multivariate-statistical treatments, or by adopting simple indicators of spatial aspects of travel, e.g. the distribution of trip lengths or distances from the city center of destination locations. On the other hand, a richer accumulation of empirical findings can be found on the association between trip-chaining behavior

and demographic and socio-economic attributes of individuals and households. For reviews, see, for example, Golob and Golob (1981) and Strathman et al. (1994).

#### 4. Classification approaches

One of the approaches that can be taken to overcome the difficulties in representing spatial patterns of travel is classification analysis, where spatial patterns are grouped into several categories based on their commonality. Kansky (1967) applies a grouping procedure to factor scores that comprise measures of travel patterns to form a seven-group classification scheme. The variables used in the factor analysis include demographic attributes (age and sex only), indicators of travel patterns (length of all trips, number of trips, average velocity, etc.), indicators of spatial patterns, and indicators of location.

Kansky's study is notable because of the elaborate measures developed to represent spatial patterns, including a "measure of circuity," defined as the average of the squared difference between the actual trip distance and the airline distance between the origin and destination; the ratio of the number of "edges" (= trips) and the number of "vertices" (= stop locations); and the ratio of the "total mileage of a travel pattern" and "the sum of local degrees of the vertices of the travel pattern," where local degree is the number of edges intersecting at a vertex. The residence location is represented by the distance from the center of the CBD, the shortest distance to the major expressway in the study area, and the shortest distance to any other expressway. Destination locations, on the other hand, are represented in terms of the circuity of a trip, linkage of destinations by trips (patterns of trip chaining), and the geometry of the travel pattern. No attributes of destination locations are incorporated, including their spatial distribution; the classification is thus not based on patterns of spatial interaction of destination locations. A similar approach can be found in Horton and Hultquist (1971), who use similar "centrographic" or "geo-statistical" measures to find groups of similar travel patterns using 1 month travel diary data from 84 households in Cedar Rapids, Iowa.

Oppenheim (1975) used cluster analysis to classify a sample of 1018 residents of the San Francisco Bay Area into 11 types of urban residents on the basis of overall similarity of personal and environmental characteristics, and independently, into nine types on the basis of their travel patterns. The variables with spatial elements are: level of public transportation service at the place of residence, represented by the distance to a bus stop, number of bus lines to work place, etc.; and "travel behavior criteria," which include the number of shopping areas patronized, frequency of shopping in the first shopping area, and distance to the first shopping

area. Measures of diversification and spatial extension of action space are thus included in the analysis in their simplest forms.

Hanson and Hanson (1981) adopt 35 day self-administered travel diary data from Uppsala, Sweden, where all stops are geo-coded to  $x$ - $y$  coordinates corresponding to street addresses. The variables of their analysis include: 17 measures of trip generation (including the number of trip chains by the number of stops); measures of "the amount of variety in the individual's activity pattern"; measures of mode use; "time spent and distances traveled for each of the five standard trip-purpose categories"; various measures of distance traveled (by day, per trip, between stops, etc.); and centrographic measures that attempt "to capture the two-dimensional nature of an individual's travel-activity pattern." Principal components analysis applied to the data produces seven factors termed "frequency of travel," "dispersion of destinations visited," "shopping, variety, and multi-stop travel," "travel to work," "social travel," "travel to recreation," and "overall distances travelled."

More recently, Pas (1984) has developed "geometrical similarity indices" based on a hierarchy of activity pattern attributes, and proposed travel pattern typologies that include five patterns of trip chaining. Recker and McNally (1985), on the other hand, apply pattern recognition theory to activity-travel pattern classification. A similar approach can be found in Joh et al. (2001). As these brief summaries of selected studies in the literature indicate, factor analytic and clustering techniques have been applied to travel pattern data with a variety of measures that represent spatial characteristics of travel patterns. The intent of these studies has been to reduce the dimensionality in the data and to produce classification schemes or sets of composite factors that are comprehensible.

Hanson and Huff (1986) address "the problems inherent in using one-day travel records for identifying homogeneous travel behavior groups." A principal component analysis, which draws on Hanson and Hanson (1981), is first applied to the 35 day observations. The analysis yields "[s]patial extent of activity pattern" as the fifth factor, which contains "distance between home and centroid of activity space," "average distance between home and destinations," "km traveled per travel day," and "proportion of stops within 1.0 km of city center." The factors from the principal component analysis themselves are not used in the cluster analysis that follows, but the former is used to select variables for the latter. None of the variables selected represent spatial elements, however.

Pas (1988) also recognizes the problem of adopting just 1 day's worth of travel data to characterizing individuals' travel patterns or classify individuals. Pas notes that "Hirsh et al. (1986) developed and estimated a dynamic model of weekly activity pattern. The results of this work indicated that ... in selecting his/her activity pattern for any period, the individual considers the activity programs already realized and those planned for later periods." It is argued that ignoring "the weekly effect on shorter cycles, such as a day, may lead to biased predictions"

(Pas, 1988). Pas thus attempts to develop a classification scheme based on weekly behavior, using "five daily travel-activity records for each of the 112 employed people in the sample, a total of 560 daily travel-activity patterns" from the Reading Activity Diary Survey. Multidimensional scaling techniques are applied to produce a similarity index, and cluster analysis to group similar patterns. Pas notes: "The results reported here are also generally consistent with those reported ... by Hanson and Huff who find that a small number of 'best days' may be used to represent most of the daily travel-activity patterns engaged in by an individual over a five week period." The inter-dependencies among daily patterns are not examined in these studies.

A number of recent studies on activity-travel patterns have adopted a variety of analytical methods, e.g. cluster analysis and discrete choice models (Goulias and Kim, 2001), association rules (Keuleers et al., 2001), data mining (Wets et al., 2000), or descriptive statistics (Gangrade et al., 2000). None of these studies addresses spatial elements, however. While Joh et al. (2001) adopt variables representing "6 locations" in their pattern recognition study, spatial implications are not clear.

On the other hand, Wallace et al. (2000) include dummy variables that indicate whether "the household is located in a center" and whether "the tour origin is located in a center" in their analysis of the propensity to chain trips. Snellen et al. (2001) evaluate the effects of neighborhood characteristics, or "urban setting" on travel patterns. "Neighborhood characteristics" are represented by "urban form," "network type, city," "network type, neighborhood," "street network type," "distance to city center," "distance to train station," "degree of urbanization, neighborhood," "degree of urbanization, city," etc. Spatial aspects of travel are represented by "average total travel distance," average trip distance by mode, and mode share expressed in terms of the fraction of distance or travel time by each mode. One of the conclusions is that "the almost constant values for the trip-tour ratio suggest that the organization of daily activity-travel patterns is highly constant across urban settings." Snellen et al. also note: "To the authors' knowledge, this is the first study that has systematically compared some key performance indicators of activity-travel patterns across cities and neighborhoods that differ in terms of urban shape and road network type." Yet, the spatial patterns are reduced to travel distance in this study as well.

## 5. Simulation approaches

As the conceptual frameworks set forth by Chapin and Hägerstrand indicate, spatial patterns evolve under a variety of motivations and a complex set of constraints. While attempts have been made to reveal characteristics of constraints and represent behaviors under them (e.g. Jacobson, 1979; Kitamura et

al., 1981), it is not at all a trivial task to address spatial patterns fully while accounting for these constraints and the linkages between trips. In the meantime, empirical findings have been accumulated on complexities of travel behavior. For example, it has been shown that the destination of a non-home-based trip is heavily influenced by the location of the home base; and, as Hemmens (1970) predicts, the travel distance and the duration of the activity at the destination are positively correlated (Kitamura et al., 1998). By formulating multinomial logit models of destination choice in which the coefficient of travel time is specified as a function of the time of day, Kitamura et al. (1998) show that the deterrence effect of travel time changes its magnitude by time of day. Obviously spatial patterns are too complex for theoretical treatment. Simulation emerges as a logical alternative for realistic representation of travel behavior.

The Chapin and Hägerstrand schools both attempted to develop simulation models to reproduce spatial patterns, although with quite different objectives. Hemmens (1970) notes that Chapin's approach is essentially a "time-budget analysis." In Ellegård et al. (1976), a "model is presented in which activities and their time requirements can be allocated to a population and its time resources. Travel requirements are then derived from assumptions concerning the daily scheduling of employment-providing and educational activities." Emphasis here is on the allocation of time, not the replication of spatial patterns. Similar emphasis on time allocation can be found in Tomlinson et al. (1973), who take the approach of entropy maximization. Location is classified in the study as "at home," "on campus," and "in town and elsewhere."

Hemmens (1970) thus attempts to develop a simulator while emphasizing the time dimension, adopting the view that "activity choice itself can only be explained in terms of the motivation, needs, wants and capabilities of the individual. The whole range of socio-economic characteristics of the individual and the family unit ... is the source form which we will seek to explain and structure the variations and patterns in activity choice. The selection of activities by an individual may be a function of his preferences, tastes, information of alternatives, habits, or financial circumstances; and most certainly is a function of his requirements for personal and household maintenance." Although no operational simulator is reported, elements of more recent activity simulation models, such as PCATS, described below, are here.

Lenntorp (1978) characterizes the approach of the Hägerstrand school as: "A major emphasis in the time geographic approach has been placed on analysing the possibilities open to the individual and not merely his actually chosen and *observed* behaviour." Given an activity program, a particular pattern of executing the program "described in time-space terms, can them be rotated among various spatial locations in a region to sense the environment by computing the number of alternative ways in which the programme can be performed from each location." The spatial resolution of Lenntorp's simulation is defined by the square grid

coordinate system "with a cell size of 250 by 250 meters," and the travel modes are "walking, cycling, car travel and travel by public transport (bus)." The development of the simulation is motivated by the recognition, "the important issue is not to make behaviour conform to some imposed norms, but rather to make room and give leeway for various possibilities of action within city-regions"; replicating spatial patterns is not the intent of the study.

Following these efforts, activity-based analysts of travel behavior have developed simulators to replicate activity-travel patterns (e.g. Recker et al., 1986a,b) in which discrete choice models of activity location drive spatial patterns. For example, in the simulator by van der Hoorn (1983), the day is divided into 96 quarter-hour periods, and individuals not committed to mandatory activities are assigned with activities at certain locations using discrete-choice models in the respective periods. A similar approach is adopted in PCATS (Kitamura and Fujii, 1998) in which the day is divided into "blocked periods" where the individual is committed to mandatory activities, and "open periods" where he or she is free to engage in discretionary activities at any location within Hägerstrand's prism.

Timmermans and his colleagues have developed a simulator, Albatross (Arentze and Timmermans, 2000), which is described as "a multi-agent rule-based system that predicts activity patterns" (Arentze et al., 2001). In Albatross, "rules rather than algebraic equations were used to represent and predict activity-travel patterns." It also incorporates the notion of "fixed activities" such as work activities or "bringing children to school." Albatross then introduces additional activities and determines "the profile and schedule position of these activities. Activity dimensions considered in profiling include travel party, duration, time-of-day, trip chaining, transport mode and location."

PCATS (Kitamura and Fujii, 1998), which stands for Prism Constrained Activity-Travel Simulator, functions similarly to Albatross, except that activity and travel are simulated using a set of discrete-choice models and duration models that determine the type, duration, location, and travel mode for each activity, and the trip associated with it. Explicitly represented in PCATS are Hägerstrand's prism constraints, and coupling constraints associated with travel modes. For example, for each choice of activity location, zones that can be reached using travel modes that are available are enumerated to define the choice set. Activity and travel are simulated along a continuous time axis; thus, trip chaining is automatically represented in PCATS.

Arentze et al. (2001) note that nested-logit models (e.g. Ettema et al., 1997; Wen and Koppelman, 1999) are the most widely applied activity-based models in transportation, which may be viewed as extensions of "tour-based models" (Gunn et al., 1987). In these models, possible travel patterns are represented by discrete alternatives, and the decision underlying observed behavior is depicted as a choice from among a set of feasible alternatives. The obvious problem that this approach presents is the enormity of potential alternatives when it is desired to represent

spatial elements in any resolution that can reasonably meet practical requirements. In fact many of these models completely disregard the spatial dimension.

Albatross, PCATS, and the discrete-choice-based “linked logit and Poisson model” (LLPM) are compared in Arentze et al. (2001). It is noted: “During the development of the LLPM, we considered and tested these alternative [nested-logit] specifications. However, we consistently found that the parameter for the logsum of the various nests was outside the required 0–1 range. The system also turned out to be highly unstable. ... Ultimately, therefore, we ended up with a set of linked logit–Poisson models that seemed to describe the observed activity–travel patterns best.” In LLPM, “[t]hus, first the number and types of out-of-home activities to perform during a particular day are modelled ... the activities are allocated to the tours. Next, for each tour, the sequencing of activities is predicted. Having modelled the organization of activities into tours, location choice is predicted next. First, the model predicts the choice of a region, and then it predicts the choice of zone within the region. ... once the destinations are known, transport mode choice is predicted.”

Predictive capabilities of the three model systems are compared by examining how each system replicates various O-D matrices for a holdout sample. A system of 19 geographical zones is used in the comparison. It is reported that, “[a]s it turns out, the performance of Albatross and PCATS [measured by the contingency coefficient] is approximately the same for each matrix. Both models outperform the LLPM, but the difference is small in the case where only [trips for] flexible activities are included” (Arentze et al., 2001).

One of the difficulties in simulating spatial patterns is the number of discrete alternatives that are involved in underlying choices, in particular when spatial representation involves a large number of geographical zones or other analytical units. When the decision underlying spatial patterns is formulated as a discrete choice whose alternatives are daily travel patterns, as in Adler and Ben-Akiva (1979) or Bowman and Ben-Akiva (2001), the choice set to be adopted for prediction can be astronomical even when the study area is represented by a relatively small number of geographical zones. Recent applications of the Markov chain Monte Carlo algorithms (e.g. Yamamoto et al., 2001) suggest that, at least for non-hierarchical logit models, this problem can be resolved quite efficiently.

Overall, efforts directed to the analysis and replication of spatial patterns in their entirety are limited; most studies and methods are trip based, whose limitations have been discussed in this chapter. The study by Arentze et al. (2001) is a rather rare case where the reproducibility of spatial patterns by models is examined with respect to O-D matrices. Whether the predictive accuracy and resolution provided by these simulators is adequate is yet to be determined in light of the purpose and context of application. When one considers how involved

spatial processes are with linked trips and the multitude of constraints governing them, it is evident that micro-simulation of spatial behavior represents a practical approach for the realistic representation of spatial processes.

## References

- Adler, T., and M. Ben-Akiva (1979) "A theoretical and empirical model of trip chaining behavior," *Transportation Research B*, 13:243–257.
- Arentze, T.A. and H.J.P. Timmermans, eds (2000) *ALBATROSS: a learning based transportation oriented simulation system*. Eindhoven: EIRASS: European Institute of Retailing and Service Studies, Eindhoven University of Technology.
- Arentze, T. A. Borgers, F. Hofman, S. Fujii, C. Joh, A. Kikuchi, R. Kitamura, H. Timmermans and P. van der Waerden (2001) "Rule-based versus utility-maximizing models of activity-travel patterns: a comparison of empirical performance," in: D. Hensher, ed., *Travel behaviour research: the leading edge*. Oxford: Elsevier.
- Beckmann, M.J. and T.F. Golob (1972) "A critique of entropy and gravity in travel forecasting," in: G.F. Newell, ed., *Traffic flow and transportation*. New York: Elsevier.
- Ben-Akiva, M. and S. Lerman (1979) "Disaggregate travel and mobility choice models and measures of accessibility," in: D.A. Hensher and P. Stopher, eds, *Behavioural travel modeling*. London: Croom Helm.
- Bowman, J. and M. Ben-Akiva (2001) "Activity based disaggregate travel demand model system with daily activity schedules," *Transportation Research A*, 35:1–28.
- Burnett, P. and S. Hanson (1979) "Rationale for an alternative mathematical approach to movement as complex human behavior," *Transportation Research Record*, 723:11–24.
- Chapin, F.S. Jr (1978) "Human time allocation in the city," in: T. Carlstein, D. Parkes and N. Thrift, eds, *Human activity and time geography, timing space and spacing time*, Vol. 2. London: Edward Arnold.
- Dijst, M. and V. Vidaković (2000) "Travel time ratio: the key factor of spatial reach," *Transportation*, 27:179–199.
- Ellegård, K., T. Hägerstrand and B. Lenntorp (1976) "Activity organization and the generation of daily travel: two future alternatives," *Economic Geography*, 57:126–152.
- Elmi, A.M., D.A. Badoe and E.J. Miller (1999) "Transferability analysis of work-trip-distribution models," *Transportation Research Record*, 1676:169–176.
- Ettema, D.F. and H.J.P. Timmermans (1997) "Theories and models of activity patterns," in: D.F. Ettema and H.J.P. Timmermans, eds, *Activity-based approaches to travel analysis*. Oxford: Pergamon.
- Ettema, D.F., A. Daly, G. de Jong and E. Kroes (1997) "Towards an applied activity-based travel demand model," in: *30th LATBR Conference*, Paper. Austin.
- Fotheringham, A.S. (1983) "A new set of spatial-interaction models: the theory of competing destination," *Environment and Planning A*, 15:15–36.
- Gangrade, S., K. Kasturirangan and R.M. Pendyala (2000) "Coast-to-coast comparison of time use and activity patterns," *Transportation Research Record*, 1718:34–42.
- Golob, T.F. (2000) "A simultaneous model of household activity participation and trip chain generation," *Transportation Research B*, 34:355–376.
- Golob, J.M. and T.F. Golob (1983) "Classification of approaches to travel-behavior analysis," in: *Travel analysis methods for the 1980s*, Special Report 201. Washington, DC: Transportation Research Board.
- Goulias, K.G. and T.-G. Kim (2001) "Multilevel analysis of activity and travel patterns: accounting for person- and household-specific observed and unobserved effects simultaneously," *Transportation Research Record*, 1752:23–31.
- Goulias, K.G., R.M. Pendyala and R. Kitamura (1990) "Practical method for the estimation of trip generation and trip chaining," *Transportation Research Record*, 1285:47–56.

- Gunn, H.F., A.I.J.M van der Hoorn and A.J. Daly (1987) "Long-term country-wide travel demand forecasts from models of individual choice," in: *5th International Conference on Travel Behaviour*, Paper. Aix-en-Provence.
- Hanson, S. (1979) "Urban-travel linkages: a review," in: D.A. Hensher and P. Stopher, eds, *Behavioral travel modelling*. London: Croom Helm.
- Hanson, S. and P. Hanson (1981) "The travel-activity patterns of urban residents: dimensions and relationships to sociodemographic characteristics," *Economic Geography*, 57:332-347.
- Hanson, S. and J.O. Huff (1986) "Classification issues in the analysis of complex travel behavior," *Transportation*, 13:271-293.
- Hanson, S. and D.F. Marble (1971) "A preliminary typology of urban travel linkages," *East Lakes Geographer*, 7:49-59.
- Hemmens, G.C. (1970) "Analysis and simulation of urban activity patterns. *Socio-economic Planning Sciences*, 4:53-66.
- Hirsh, M., J.N. Prashker and M.E. Ben-Akiva (1986) "Dynamic model of weekly activity pattern," *Transportation Science*, 20:24-36.
- Horton, F.E. and J.F. Hultquist (1971) "Urban household travel patterns: definition and relationship to household characteristics," *East Lakes Geographer*, 7:37-48.
- Horton, F.E. and W.E. Wagner (1969) "A Markovian analysis of urban travel behavior: pattern response by socioeconomic-occupational groups," *Highway Research Record*, 283:19-29.
- Jacobson, J. (1979) "Models of non-work activity duration," Ph.D. thesis. Cambridge: Department of Civil Engineering, MIT.
- Joh, C.-H., T.A. Arentze and H.P. Timmermans (2001) "Pattern recognition in complex activity travel patterns: comparison of Euclidean distance, signal-processing theoretical, and multidimensional sequence alignment methods," *Transportation Research Record*, 1752:16-31.
- Jones, P., F. Koppelman and J.P. Orfueil (1990) "Activity analysis: state-of-the-art and future directions," in: P. Jones, ed., *Developments in dynamic and activity-based approaches to travel analysis*. Aldershot: Gower.
- Jou, R.-C. and H.S. Mahmassani (1997) "Comaprative analysis of day-to-day trip-chaining behavior of urban commuters in two cities," *Transportation Research Record*, 1607:163-170.
- Kansky, K.J. (1967) "Travel patterns of urban residents," *Transportation Science*, 1:261-285.
- Keuleers, B., G. Wets, T. Arentze and H. Timmermans (2001) "Association rules in identification of spatial-temporal patterns in multiday activity diary data," *Transportation Research Record*, 1752:32-37.
- Kitamura, R. (1983) "A sequential, history dependent approach to trip chaining behavior," *Transportation Research Record*, 944:13-22.
- Kitamura, R. (1984) "Incorporating trip chaining into analysis of destination choice," *Transportation Research B*, 18:67-81.
- Kitamura, R. and S. Fujii (1998) "Two computational process models of activity-travel behavior," in: T. Gärling, T. Laitila and K. Westin, eds, *Theoretical foundations of travel choice modelling*. Oxford: Pergamon.
- Kitamura, R., L.P. Kostyniuk, and M.J. Uyeno (1981). Basic properties of urban time-space paths: empirical tests," *Transportation Research Record*, 794:8-19.
- Kitamura, R., C. Chen and R.M. Pendala (1997) "Generation of synthetic daily activity-travel patterns," *Transportation Research Record*, 1607:154-162.
- Kitamura, R., C. Chen and R. Narayanan (1998) "The effects of time of day, activity duration and home location on travelers' destination choice behavior," *Transportation Research Record*, 1645:76-81.
- Kofoed, J. (1970) "Person movement research: a discussion of concepts," *Papers of the Regional Science Association*, 24:141-155.
- Kondo, K. (1974) "Estimation of person trip patterns and modal split," in: D.J. Buckley, ed., *Transportation and traffic theory*. New York: Elsevier.
- Lenntorp, B. (1978) "A time-geographic simulation model of individual activity programmes," in: T. Carlstein, D. Parkes and N. Thrift, eds, *Human activity and time geography, timing space and spacing time*, Vol. 2. London: Edward Arnold.
- Lerman, S.R. (1976) "Location, housing, automobile ownership, and mode to work: a joint choice model," *Transportation Research Record*, 610:6-11.

- Lerman, S.R. (1979) "The use of disaggregate choice models in semi-Markov process models of trip chaining behavior," *Transportation Science*, 13:273–291.
- Marble, D.F. (1967) "A theoretical exploration of individual travel behavior," in: W.L. Garrison and D.F. Marble, eds, *Quantitative geography, Part I: economic and cultural topics. Studies in geography*, No. 13. Evanston: Department of Geography, Northwestern University.
- Marble, D.F. and S.R. Bowlby (1968) "Shopping alternatives and recurrent travel patterns," in: F. Norton, ed., *Geographic studies of urban transportation and network analysis. Studies in geography*, No. 16. Evanston: Department of Geography, Northwestern University.
- Mikkonen, K. and M. Luoma (1999) "The parameters of the gravity model are changing – how and why?" *Journal of Transport Geography*, 7:277–283.
- Nystuen, J.D. (1967) "A theory and simulation of intraurban travel," in: W.L. Garrison and D.F. Marble, eds, *Quantitative geography, Part I: economic and cultural topics. Studies in geography*, No. 13. Evanston: Department of Geography, Northwestern University.
- O'Kelly, M.E. (1981). A model of the demand for retail facilities, incorporating multistop, multipurpose trips," *Geographical Analysis*, 13:134–148.
- Oppenheim, N. (1975) "A typological approach to individual urban travel behavior prediction," *Environment and Planning A*, 7:141–152.
- Pas, E.I. (1984) "The effect of selected sociodemographic characteristics on daily travel-activity behavior," *Environment and Planning A*, 16:571–581.
- Pas, E.I. (1988) "Weekly travel-activity behavior," *Transportation*, 15:89–109.
- Recker, W.W. and M.G. McNally (1985) "Travel/activity analysis: pattern recognition, classification and interpretation," *Transportation Research A*, 19:279–296.
- Recker, W.W., M.G. McNally, and G.S. Root (1986a) "A model of complex travel behavior: Part I – theoretical development," *Transportation Research A*, 20:307–318.
- Recker, W.W., M.G. McNally, and G.S. Root (1986b) "A model of complex travel behavior: Part II – an operational model," *Transportation Research A*, 20:319–330.
- Snellen, D., T. Arentze, A. Borgers and H. Timmermans (2001) "Impact of urban setting on activity-travel patterns: comparison of performance indicators with quasi-experimental design data," *Transportation Research Record*, 1780:1–8.
- Strathman, J.G., K.J. Dueker, and J.S. Davis (1994) "Effects of household structure and selected travel characteristics on trip chaining," *Transportation*, 21:23–45.
- Thill, J.-C. (1985) "Demand in space and multipurpose shopping: a theoretical approach," *Geographical Analysis*, 17:114–129.
- Thill, J.-C. and I. Thomas (1987) "Towards conceptualising trip chaining behaviour: a review," *Geographical Analysis*, 19:1–17.
- Timmermans, H., P. van der Waerden, M. Alves, J. Polak, S. Ellis, A.S. Harvey, S. Kurose and R. Zandee (2003) "Spatial context and the complexity of daily travel patterns: an international comparison," *Journal of Transport Geography*, 11:37–46.
- Tomlinson, J., N. Bullock, P. Dickens, P. Steadman and E. Taylor (1973) "A model of students' daily activity patterns," *Environment and Planning*, 5:231–266.
- van der Hoorn, T. (1983) "Development of an activity model using a one-week activity-diary data base," in: S. Carpenter and P. Jones, eds, *Recent advances in travel demand analysis*. Aldershot: Gower.
- Wachs, M. and J.G. Koenig (1979) "Behavioural modelling, accessibility and travel need," in: D.A. Hensher and P.R. Stopher, eds, *Behavioural travel modelling*. London: Croom Helm.
- Wallace, B., J. Barnes and G.S. Rutherford (2000) "Evaluating the effects of traveler and trip characteristics on trip chaining, with implications for transportation demand management strategies," *Transportation Research Record*, 1718:97–106.
- Wen, C.-H. and F.S. Koppelman (1999) "Integrated model system of stop generation and tour formation for analysis of activity and travel patterns," *Transportation Research Record*, 1676: 136–144.
- Wets, G., K. Vanhoof, T. Arentze and H. Timmermans (2000) "Identifying decision structures underlying activity patterns: an exploration of data mining algorithms," *Transportation Research Record*, 1718:1–9.
- Wheeler, J.O. (1972) "Trip purpose and urban activity linkages," *Annals of the Association of American Geographers*, 62:641–654.

- Wilson, A.G. (1970) *Entropy in urban and regional modelling*. London: Pion.
- Yamamoto, T., R. Kitamura and K. Kishizawa (2001) “Sampling alternatives from colossal choice set: application of Markov chain Monte Carlo algorithm,” *Transportation Research Record*, 1752:53–61.

## *Chapter 30*

# MENTAL MAPS

LISA WESTON

*University of Texas, Austin, TX*

SUSAN HANDY

*University of California, Davis, CA*

## **1. Introduction**

How do people keep track of the environment around them? How do they remember where specific things are in space and then figure out how to get “there” from “here”? Carrying around and referring to a cartographic map is one way people learn about a new environment, but not how they are likely to interact with the environment once they have some level of experience with it. Elements of the cartographic map and the environment as the traveler experiences it accrete to form a remembered map carried everywhere within the mind. Owing to different levels of interaction with the built environment and different levels of interest in learning about the built environment, the remembered maps of the residents of an area vary from one resident to another. This map, which may comprise words and feelings, as well as images, is commonly called a mental map. Mental maps are the way that people organize large amounts of data about the physical environment into a form that allows for future referral. These maps contain information about specific elements of the physical environment and their spatial relationship.

The concept of mental maps, also called cognitive maps (see Chapter 28), is a complement to the concept of spatial cognition. Research into spatial cognition concentrates on the process of acquiring knowledge about space – how people acquire knowledge about space and where in the brain that information is stored for later retrieval (Lloyd, 1999). Much of the work on this concept comes from the field of environmental psychology (e.g. Gärling, 1999). Research on the concept of mental maps concentrates on what people do with the knowledge of space they have acquired – how people use that knowledge to navigate space and the accuracy and sufficiency of that knowledge. The mental map, or collection of spatial information that one carries in their mind, is also called a cognitive map. Much of the work in this area comes from the field of geography (e.g. Allen,

1999a). The closeness of these two concepts means that the research from each field often overlaps. Most notably, researchers in both fields have explored the concept of wayfinding, the way that people move successfully through the physical environment to reach a desired destination. Wayfinding and its implications for urban designers and planners are the focus of this chapter.

Mental maps provide a basis for wayfinding and for solving the spatial problems associated with wayfinding. These spatial problems include choosing a destination, determining a route between two points, choosing an alternative route when the primary route is impassable, navigating along a route, and learning a new spatial environment. Some researchers further differentiate among regular travel between two known points (commute), irregular travel from a known point to an unknown point (quest), and loop travel starting at a known point and journeying through unknown areas and arriving back at the starting point (explore) (Allen, 1999b).

Whatever the travel purpose, individuals will evaluate the environment around them to either verify existing information or note new features. These elements of the physical environment have characteristics that different individuals will assess in a relatively consistent way (color, shape, size, smell, etc.), and they may have particular associations that vary from person to person (location of one's first date, car trouble, workplace, etc.). Because individuals will prioritize characteristics and associations according to personal experience, elements of the built environment and their importance will vary in the mental maps of different individuals. As a result, the solution of spatial problems will vary from person to person.

Design professionals are interested in mental maps because they give an indication of how "readable" a city is. Do the elements of the built environment work together to facilitate navigation of the city? Are there places where people are consistently confused? By understanding how the parts of the built environment interact to form images in people's minds, design professionals can propose changes as necessary to make the built environment easier to comprehend for residents and visitors alike. An understanding of mental maps can also help to improve navigation media as well as emergency search procedures (Kitchin and Freundschuh, 2000).

Transportation professionals are interested in how people learn, remember, and respond to the physical environment because it affects the choices they make about travel. First, a resident's mental map of her community reflects her knowledge of potential destinations and routes and thus the alternatives she considers when making her choice of destination or route. Second, her mental map influences her perceptions of the distances to potential destinations by different routes and modes and thus her valuation of (or utility for) each possibility. In addition, her mental map may include associations that make certain choices more or less attractive than other choices. She may consider one

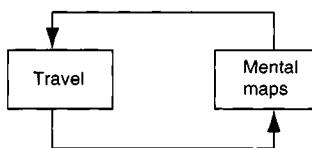


Figure 1. The relationship between travel and mental maps.

route more attractive than other or perceive a particular route to be unsafe for walking. Understanding mental maps can thus provide important insights into travel behavior.

Furthermore, understanding differences in mental maps can provide important insights into differences in travel behavior. In efforts to understand the travel behavior of particular user groups such as children and the elderly, an important exploratory step is to identify differences in how individuals perceive and remember elements of the physical environment and how well those remembered elements correlate with the physical environment. An element of the transportation system such as a highway may be perceived as a facilitator of travel by someone with a car but as a barrier to someone without – a disabled person, an elderly person, or a child.

Travel is intimately tied to mental maps. Through travel, people acquire the spatial knowledge that makes up their mental maps (Figure 1). Their mental maps then shape their choice about travel. Understanding mental maps is thus important for understanding and for modeling travel behavior, as described more fully in Chapter 28. With the goal of introducing transportation planners to the concept of mental maps, this chapter looks at the following questions:

- What are mental maps?
- How do people create mental maps?
- How have mental maps been used?
- How can transportation professionals use mental maps?

## 2. What are mental maps?

Mental maps are representations of spatial knowledge in our memories (Gärling and Golledge, 2000). They are not like cartographic maps. They are not static, and they may include more than the two-dimensional depiction of roads, rivers, and other physical features typically associated with cartographic maps. Strictly speaking, mental maps contain things people associate with a particular physical environment. These characteristics of the physical environment may include visual attributes such as the type of element (a store, road, river, etc.), non-visual

attributes such as sounds and smells, and non-physical attributes of physical locations (hours of operation, historical meanings, personal feelings, etc.). One way to think about mental maps is as a collage of information (Tversky, 2000).

A basic mental map created by a person shows how they perceive the relative location of known elements in the area. Depending on the person's familiarity and experience with the area, the map will be distorted in different ways from the physical reality. A classic example of a distorted mental map is the widely reproduced cover of a past issue of *The New Yorker* magazine that shows the New Yorker's view of the USA. The foreground consists of a detailed map of 9th And 10th Avenues, beyond the Hudson River is a small strip labeled "Jersey," and beyond that is a strip with labels denoting places such as Chicago, Kansas City, Las Vegas, Nevada, and Texas. In the background is the Pacific Ocean and Asia. While New York itself is rich and detailed, the rest of the nation, and indeed the world, is an amorphous place with few distinguishing characteristics.

Although this is an extreme example, everyone has a distorted view of the city around them to some degree. Work by Golledge (1999) looks specifically at how people distort maps of cities. A newcomer to a city may have an irregularly distorted map that may come to more closely resemble reality the more familiar they become with their new environment. Over time, potentially confusing elements of the urban environment, such as streets that are parallel for some distance and then cross each other, may be resolved. So, unlike cartographic maps that change slowly as new features are built on the landscape, mental maps are continuously flexible as people learn new parts of the built environment.

### **3. How do people create mental maps?**

Although cartographic maps provide important spatial information to travelers, particularly about the straight-line distances between points, direct experience with the environment is the primary and most accurate source of spatial information for mental maps (Gärling and Golledge, 2000; Kitchin and Freundschuh, 2000; Tversky, 2000). A person acquires knowledge of a place when moving through it, and builds a mental map based on this knowledge. Four main stages in the process of creating a mental map appear in the writings of several authors (Downs and Stea, 1977; Gärling et al., 1997; Golledge and Stimson, 1997): first, a person notices some element in the built environment, then assigns characteristics to this element, then stores this data point for future use, and, finally, when needed, retrieves the information. Each of these steps will be considered in turn.

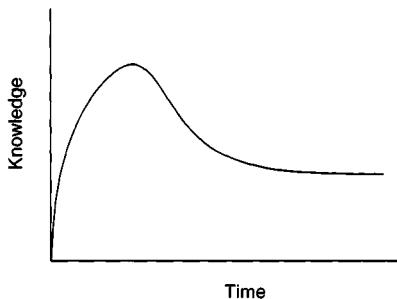
What do people notice in the built environment? This question is often asked by marketing and advertising specialists, as well as urban designers. Attracting the attention of people passing by is the first step in turning them into customers.

Visible and well-known symbols that are easily noticed by passing drivers have proven an effective business strategy. Urban designers are interested in clues from the landscape that give people information about where they are, whether downtown, a residential neighborhood, or an art district, for example. Although urban designers are ultimately aiming at a quality beyond just the physical parts of the urban environment, the parts themselves are important. Because of the overwhelming amount of information flooding the senses in the urban environment, people only “see” and remember select parts of the built environment. These may be parts that stand out due to their arrangement in the panorama: a building quite a bit taller than those nearby, a building, statue, or other object at the end of a major street, for example.

Next, people assign attributes to the element. Many researchers have addressed this step, but the work by Lynch (1960) is arguably the most well known. Lynch asked a sample of people in Boston, Jersey City, and Los Angeles to draw maps of their cities. He and his colleagues then analyzed the results, looking for commonalities and differences among the maps. Based on this analysis, Lynch grouped the elements of the built environment into five types: paths, edges, nodes, districts, and landmarks. Paths and edges are linear elements. While paths characterize a way to travel through the environment, edges are barriers or delineations of the end of one type of space and the beginning of another. Nodes and landmarks are points in the physical environment. A landmark is a typically large element in the landscape that is easily distinguishable. A node is a point that a person enters. It is characterized as the concentration, or foci, of the physical environment. It is often a place where the traveler must make a decision. Some elements may fall into more than one category, depending on the view of the observer. For example, an elevated highway may be a path to a driver and an edge to a pedestrian. Although all the elements are important parts of the built environment, paths have a role separate from the others. It is via paths that people interact with the built environment and learn their way around a city.

These five types of element in the built environment make up what Downs and Stea (1977) describe as equivalence categories. In equivalence categories, elements are grouped according to their similarities. Sameness is emphasized and differences are glossed over. For example, the equivalence category of paths includes the Pacific Coast Highway, Route 66, Main Street, and the Appalachian Trail. The other way to categorize elements is by emphasizing differences and glossing over similarities, or what are called identity categories. The identity category of Santa Fe in New Mexico may include the plaza, St Francis' Cathedral, Cerrillos Road, piñon trees, and many other elements, big and small.

Once a person assigns characteristics to an element, it is stored for future use. The neuroscience literature examines the physiological aspect of where in the brain people store information about space and how they manipulate it. This aspect of storing spatial information will not be covered here. Instead, one can



**Figure 2.** Possible change in spatial knowledge over time.

simply take the elements as being stored in a person's mental map. Items are added, subtracted, and otherwise altered as a person gathers more information about an area.

At some later point a person retrieves the data. At this time, the element may be remembered inaccurately, further adding to inaccuracies that may have been included when the individual first stored the item. As the person assembles stored data about the built environment for the task at hand, the reconstructed map may not come back together the way it came apart. Distortions in mental maps commonly include distance biases and simplifications. People tend to overestimate short distances and underestimate long distances (Gärling and Golledge, 2000; Tversky, 2000). They tend to overestimate distances when a route includes barriers, clutter, turns, a large number of nodes, and when they retain a large amount of information from the environment; an aesthetically pleasing environment leads to underestimates of distances (Tversky, 2000). Although a mental map becomes more accurate as a person's experience with a place increases, her knowledge of a place may actually decline over time as their experiences become limited to certain habitual paths and they stop acquiring new information about the place (Gärling and Golledge, 2000), as depicted in Figure 2.

A person may reconstruct a map for one of several purposes. In general, these uses have to do with formulating and executing travel plans (Gärling and Golledge, 2000). One purpose is for identifying possible destinations and routes from which a person can choose. Another purpose is for wayfinding. As a person travels from one place to another she monitors her progress through space to assure that she is on the right path. This process may consist of checking off certain features that indicate which path she is on. A third purpose for one's mental map is to give information about the urban environment to another person. A person may draw a map for an out-of-town guest or give verbal instructions over the phone. The many elements of one's mental map must be translated into physical images or words or both to convey the spatial information stored in it to others.

Figures 3–6 are drawings of four mental maps of Austin, Texas, that demonstrate the wide variability in mental maps that different individuals formulate for the same place based on their different experiences. Figure 3 was drawn by an 11-year-old girl, who drew downtown as an amorphous cluster of high rises and included her preschool and her favorite restaurant. What is not on the map is also interesting: her home and present school are not to be found. Figures 4–6 were drawn by three undergraduate students ranging in age from 22 years to 26 years and having lived in the area from 5 years to 10 years. Provided with the same directions for drawing a mental map, these are the maps they provided. The author of Figure 4 relies on landmarks and districts to characterize Austin, and only two major paths are included. Clearly, when asked to draw his mental map of Austin, this author focused on physical images rather than spatial relationships. In contrast, the author of Figure 5 concentrates more on districts and paths with very few landmarks. The author of Figure 6 instead provided a picture of Austin within the larger context of the surrounding Hill Country. Note how dramatically these figures vary in their emphasis on different types of elements (landmarks, districts, paths, etc.) and in the examples of each element on which they focus (e.g. a favorite restaurant or the capitol as a landmark). They reflect differences in the experiences of the authors and the spatial knowledge they retain.

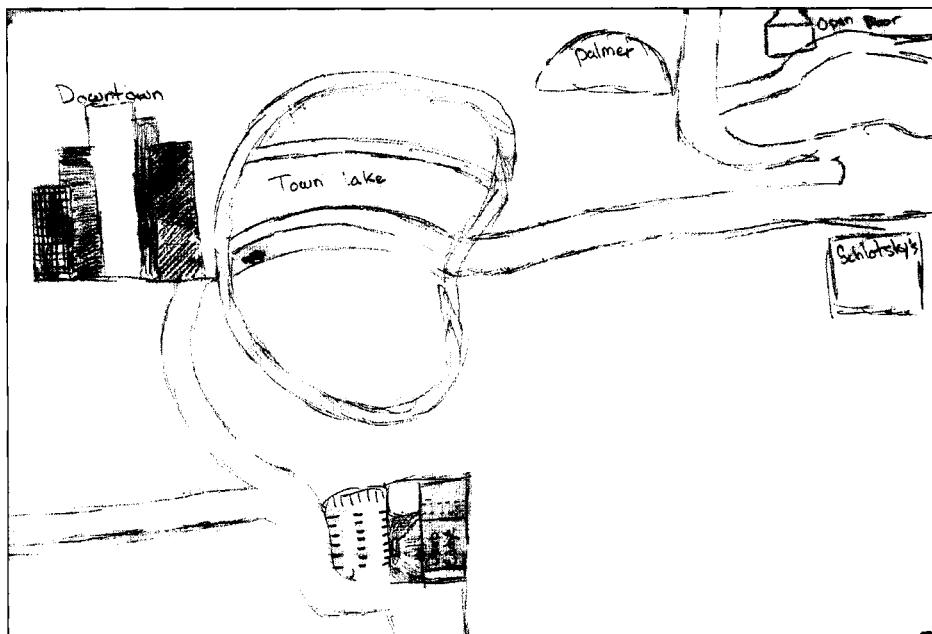


Figure 3. Representation of mental map of Austin, Texas: 11-year-old girl.

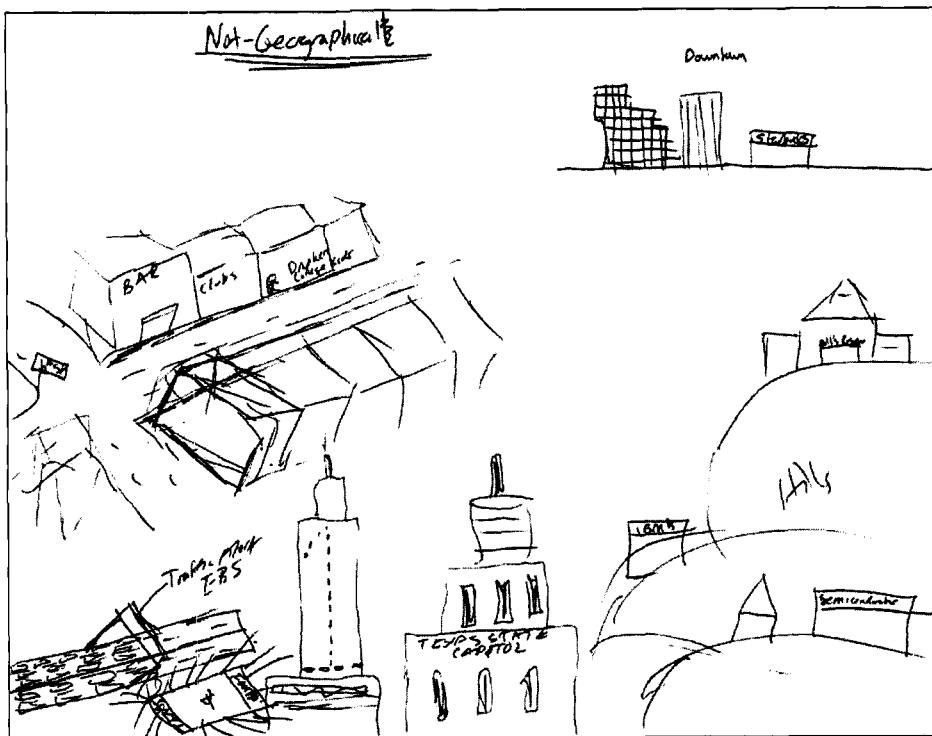


Figure 4. Representation of mental map of Austin, Texas: first young man.

#### **4. How have mental maps been used?**

Researchers have used drawings of mental maps in their efforts to understand spatial cognition. But such drawings have important practical uses as well. It is possible to assemble the mental maps of a sample of individuals to create an aggregate map that gives a sense of both the average understanding of the environment and the variation in that understanding. Although such drawings cannot be superimposed in a simple way, given their scale differences and spatial distortions, the specific elements included can be identified and tallied for a set of maps. Such information can then be recorded on a simple base map of the area to create an aggregate mental map. The advent of geographic information systems may provide new ways of both collecting and analyzing data from mental maps and expand the range of practical applications for mental maps. Mental map projects have been used to assess the importance people assign to places (US Environmental Protection Agency, 2002), identify areas of their city they consider

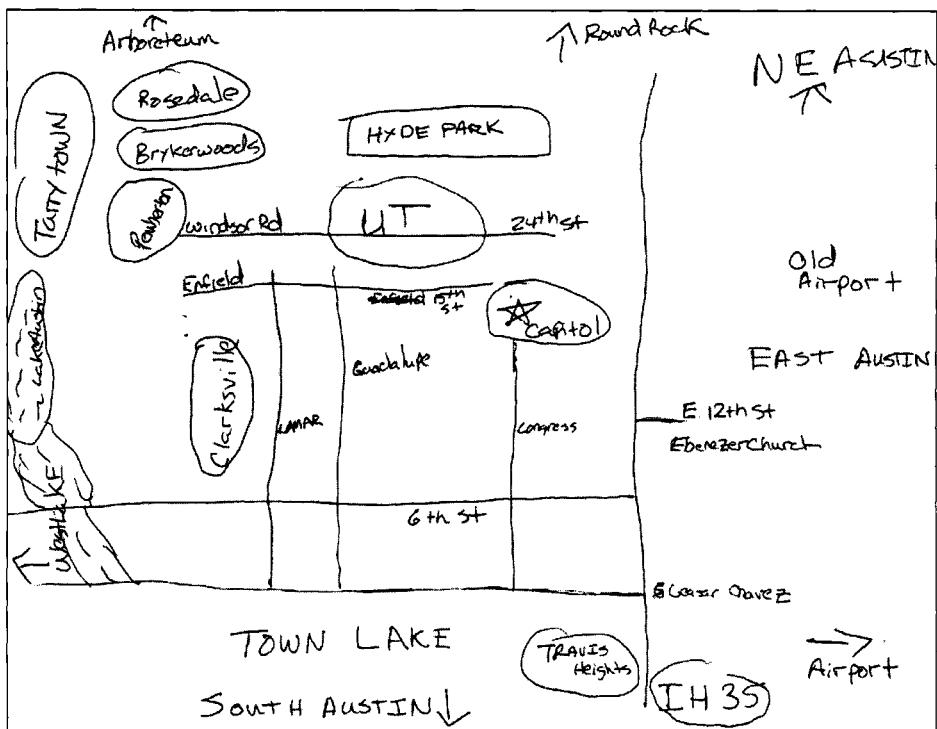


Figure 5. Representation of mental map of Austin, Texas: second young man.

dangerous (Matei et al., 2001), or gain knowledge from special interest groups about their assessment of their surroundings (Halseth and Doddridge, 2000).

The US Environmental Protection Agency (2002) has prepared a guide to help communities plan for community-based environmental protection efforts. Preparation of what they call cognitive maps is one of several recommended methods to assess how people feel about their community. Working from the assumption that cultural behaviors have created certain environmental problems, they recommend that local groups first assess this cultural picture of a community in order to identify potential solutions to the problem. The cognitive map process offers a way to identify what is important to the members of the community, what landmarks or features are associated with the community, and create an image of the community that incorporates intangible aspects as well as physical elements.

Another intangible aspect of city life that can be identified through the mental mapping process is the identification of areas where people feel danger. Matei et al. (2001) compiled the results of 215 mental maps of the southern third of Los Angeles County. One hypothesis that the researchers tested was the influence of

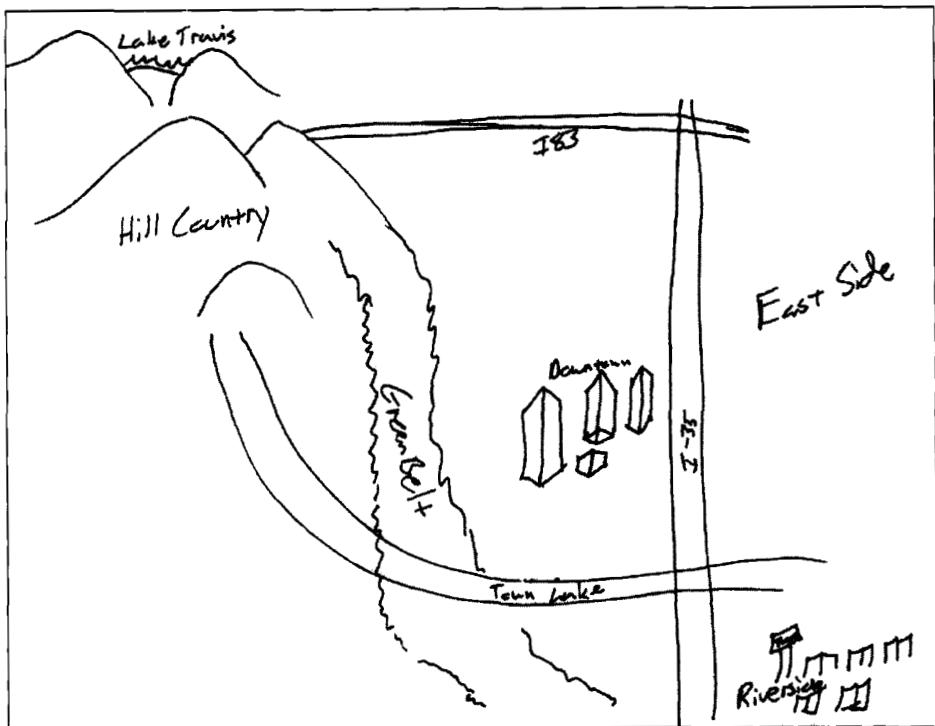


Figure 6. Representation of mental map of Austin, Texas: third young man.

communication media on people's comfort maps – where they felt safe and where they did not. They found that television coupled with personal interactions produced distortions in mental maps, where distortion was defined as the perception of crime victimization compared with actual crime statistics. Comparing mental maps with crime data provides urban policymakers with raw data to build policy on. Similar work has been done in other US cities (Ley, 1972) to identify areas where residents are fearful.

While Matei et al. (2001) worked with adults who can articulate their perceptions of the urban environment, Halseth and Doddridge (2000) have developed a mental mapping tool that allows children to express their preferences about the urban environment. Their tool is designed as a regular curriculum element. While susceptible to the instructor's influence, they found that mental mapping exercises can be effective in understanding how children perceive the urban environment. This is the first step in including the needs of children in urban planning.

These three examples show the breadth of the use in mental maps to allow people to provide a physical representation of intangibles that affect their

interaction with the urban environment. It also allows for the inclusion of young people and people with different backgrounds.

## 5. How can transportation professionals use mental maps?

As suggested in the examples above, mental maps can be used as a tool to aggregate impressions about a city from a variety of people. Lynch (1960) identified several aspects of the transportation network that made Boston a difficult city to read and consequently navigate, five-sided “squares” and streets with multiple names being two of the most obvious. This information can then be used to guide improvements to benefit travelers. Communities experiencing a high crash rate in a particular area might gain from asking residents to describe the area, either in words or images. The consistent overlooking of important directional information may indicate that it is not prominent enough to aid travelers. Subway map makers face the challenge of producing a map that accurately portrays the various lines and the parts of city they access, but does not distort the city to the extent that a traveler cannot reach the desired destination.

The principles of mental maps are not confined to post-mortems of transportation infrastructure but can also be applied to the building of new infrastructure. Lynch (1960) proposes several design guidelines for paths to make cities more readable. First, the main streets of a city should have a distinctive characteristic so that travelers know they are on a main thoroughfare. Potential distinctive treatments include planting one species of tree consistently along the road or in the median, locating unique businesses along the path, or installing a special lighting treatment. Second, the direction of the path should be clear. Several sharp turns or a series of gentle curves that maintains the general direction of the path all help travelers keep a sense of direction, while gentle curves that lead to a major shift in direction can be confusing. Third, some sort of gradation that denotes direction is useful to travelers. Some of Lynch's suggestions are gradient, increased density of signs near a shopping node, change in block length, or a distinguishable feature to one side – such as a park. Fourth, a series of local landmarks assists travelers in noting their passage along the path. These small landmarks also serve as a way to locate smaller elements along the path. Fifth, an overall visual sensation, for example a preview of a major destination such as downtown, impresses travelers with a heightened image of a city. Passing through or next to distinctive districts also helps the traveler orient herself in the city.

The last two elements that Lynch recommends for paths to assist in the readability of cities are related to intersections and the overall feel of the path. Intersections are potential decision points for travelers; therefore the ability to convey information about the city in an easily understood way is especially imperative. They should be easy to navigate and give a sense of how a city is laid

out. Besides having regular, gridiron intersections, Manhattan has long blocks in one direction and short blocks in the other. This helps travelers easily determine if they are going north-south or east-west. The gridiron overlaid with radiating street pattern of Washington, DC, produces some intersections that are geometrically difficult to navigate. However, the terminating monuments and buildings help in determining direction. Consideration of the person navigating the paths of a city is the motivating force behind Lynch's recommendations. Therefore, his final recommendation regarding paths through urban areas is that there be an overall organization to a major path that he describes as "melodic." This overall organization would provide variety to the traveler. Instead of being presented with sameness along a path, for example, all tall buildings, or a straight journey, the traveler would encounter a variety that would engage her and teach her about the area. These principles were applied in a theoretical way to Boston's Central Artery (Appleyard et al., 1964). Before the highway was originally built, Appleyard et al. suggested a route that would engage and inform the traveler, in addition to providing the through route the transportation engineers were seeking. Ironically, the Central Artery has recently been removed, partially in the interests of reducing the physical and mental barrier it created. This change will undoubtedly lead to a significant revision to the mental maps of Boston residents.

## 6. Conclusions

Mental mapping exercises are a promising yet underutilized tool for the transportation field. For travel behavior researchers, mental maps can provide important insights into the choices that travelers perceive to be available to them and the ways that they evaluate those choices. For transportation planners, mental maps can provide a way of identifying critical paths and perceived barriers that may help in targeting planning efforts, prioritizing capital investments, and designing community-sensitive facilities. Although often used to evaluate existing conditions, previous work with mental maps suggests several principles for the design of transportation infrastructure that engages and informs the traveler. Because paths are the primary way that people interact with and learn about the city and because they are a major element in the urban fabric, attention to how paths may help or hinder the ability of people to navigate the city is warranted. Mental maps provide an important tool for transportation professionals to use toward that end.

## References

- Allen, G.L. (1999a) "Spatial abilities, cognitive maps, and wayfinding: bases for individual differences in spatial cognition and behavior," in: R.G. Golledge, ed., *Wayfinding behavior: cognitive mapping and other spatial processes*. Baltimore: The Johns Hopkins University Press.

- Allen, G.L. (1999b) "Cognitive abilities in the service of wayfinding: a functional approach," *Professional Geographer*, 51:554–561.
- Appleyard, D., K. Lynch and J.R. Myer (1964) *The view from the road*. Cambridge: MIT Press.
- Downs, R.M. and D. Stea (1977) *Maps in minds: reflections on cognitive mapping*. New York: Harper and Row.
- Gärling, T. (1999) "Human information processing in sequential spatial choice," in: R.G. Golledge, ed., *Wayfinding behavior: cognitive mapping and other spatial processes*. Baltimore: Johns Hopkins University Press.
- Gärling, T. and R.G. Golledge (2000) "Cognitive mapping and spatial decision making," in: R.M. Kitchin and S. Freundschuh, eds, *Cognitive mapping: past, present and future*. London: Routledge.
- Gärling, T., M. Selart and A. Book (1997) "Investigating spatial choice and navigation in large-scale environments," in: N. Foreman and R. Gillett, eds, *Handbook of spatial research paradigms and methodologies*, Vol. 1. *Spatial cognition in the child and adult*. Hove: Psychology Press.
- Golledge, R.G. (1999) "Human wayfinding and cognitive maps," in: R.G. Golledge, ed., *Wayfinding behavior: cognitive mapping and other spatial processes*. Baltimore: Johns Hopkins University Press.
- Golledge, R.G. and R.J. Stimson (1997) *Spatial behavior: a geographical perspective*. New York: Guilford Press.
- Halseth, G. and J. Doddridge (2000) "Children's cognitive mapping: a potential tool for neighbourhood planning," *Environment and Planning B*, 27:565–582.
- Kitchin, R. and S. Freundschuh (2000) "Cognitive mapping," in: R.M. Kitchin and S. Freundschuh, eds, *Cognitive mapping: past, present and future*. London: Routledge.
- Ley, D. (1972) "The black inner city as a frontier outpost: images and behavior of North Philadelphia Neighborhood," Ph.D. thesis. Pennsylvania: University Park.
- Lloyd, R. (1999) "Organization of feature-, time-, or location-based mental models," *Professional Geographer*, 51:525–538.
- Lynch, K. (1960) *The image of the city*. Cambridge: MIT Press.
- Matei, S., S.J. Ball-Rokeach, and J.L. Qiu (2001) "Fear and misrepresentation of Los Angeles urban space: a spatial-statistical study of communication-shaped mental maps," *Communications Research*, 28:429–463.
- Tversky, B. (2000) "Levels and structure of spatial knowledge," in: R.M. Kitchin and S. Freundschuh, eds, *Cognitive mapping: past, present and future*. London: Routledge.
- US Environmental Protection Agency (2002) *Community culture and the environment: a guide to understanding a sense of place*. Washington, DC: US EPA, Office of Water.

***Part 8***

**GEOSIMULATION**

## GEOSIMULATION, AUTOMATA, AND TRAFFIC MODELING

PAUL M. TORRENS

*University of Utah, Salt Lake City, UT*

### 1. Introduction

Recent developments in the research landscape have made possible a new paradigm for spatial simulation, what is coming to be known in the geographical sciences as the geosimulation approach. This novel approach to simulation development is characterized by detailed, dynamic, and interactive simulation environments, often operating in near real time and exhibiting very realistic characteristics. A new class of “microscopic” simulation has begun to emerge around the approach, focused on automata-based tools for model building (Torrens, 2002). This chapter discusses the potential of geosimulation for traffic modeling, and describes how geosimulation-style tools – cellular automata (CA) and multi-agent systems (MAS) – have been used to build a variety of vehicle and pedestrian traffic simulations. The chapter also explores some of the current limitations of the field, particularly as an applied science, and discusses future avenues of potential research inquiry.

### 2. Recent developments in the research landscape

The emergence of a new class of simulation tools for traffic modeling has been catalyzed by recent developments in several fields, including computer hardware, computer science, “dataware,” and complexity studies.

The computer hardware now available for running transport simulations is unprecedented when compared with that available only a few years ago. Advances in central processing units, graphics processing units, data storage, bandwidth, and parallel computing have made possible the construction of highly detailed and dynamic simulation environments for studying traffic, and in many cases these models can be run from desktop machines.

In parallel, important advances in computer science have influenced traffic simulation, either by providing new programming environments for developing

simulation software or by introducing new methodologies for formulating traffic models. Object-oriented (OO) programming languages – Java and C# are popular examples – have been particularly influential. They have provided a new knowledge representation paradigm for applied science (Gimblett, 2002). The OO approach is particularly useful for simulation because it provides an intuitive framework for binding entities (objects) and the behavior (methods) that governs their interactions. Other developments in computer science, such as artificial intelligence (Kurzweil, 1990, 1999) and artificial life (Levy, 1992), have also been influential, particularly in introducing new simulation techniques. Artificial neural networks (Gurney, 1997) have been particularly influential, as have been methodologies borrowed from intelligent agents research (Schleifer, 2002). Automata-based tools (Sipper, 1997) have been especially significant; they form the basis for geosimulation-style tools in most of the examples that will be discussed in this chapter.

In many instances, recent advances in transport simulation have been supported by developments in the “dataware” used to support modeling. Advances in geographic information science have provided geographic information systems for storing, manipulating, and visualizing spatial data used in building models, and spatial analysis has provided new methodologies for processing that data. A new field, geographic information systems for transportation (GIS-T), has emerged in recent years (Thill, 2000; Miller and Shaw, 2001). Advances in photogrammetric and geomatic engineering have also provided new, remotely sensed, data sources for transport models.

Transport model developers are also finding new ways to interpret – and model – transport systems as complex adaptive systems, using ideas from complexity studies. The idea of emergence (Holland, 1998), which characterizes systems as the product of bottom-up and local scale interactions between independent components, has been widely adopted. This replaces reductionist approaches, which treat systems as simple top-down aggregations of system parts.

Ultimately, these developments have had important implications for the ways in which we now model transport systems. New opportunities for developing models with hitherto unseen degrees of realism are now available. In particular, these developments have facilitated advances in the representation of dynamics, scale, interaction, and entities in transport models, broadening the range of systems that modelers can now simulate, as well as revitalizing the ways in which we consider simulation as an exercise.

### **3. The emerging geosimulation approach**

Geosimulation is a catch-all phrase that can be used to represent a “new wave” of simulation that has come to prominence in recent years. The geosimulation approach builds on advances that have been discussed above, drawing together

a diversity of theories and techniques across disciplinary boundaries, offering unique and innovative perspectives on spatial simulation. The approach is used most prominently in urban simulation, and has also been widely used to build traffic models.

Geosimulation-style traffic models have a number of innovative characteristics that distinguish them from “traditional” approaches, and these attributes draw, in most cases, directly from the advances that were discussed in the last section.

The first distinguishing aspect relates to the depiction of time. In many cases, traffic models designed in the geosimulation framework operate in near real time, with time divided into discrete “packets of change” (Anderson, 2002) that approximate the reaction time of drivers or pedestrians. Geosimulation-style traffic models are often dynamic in other senses, with simulated entities reacting to evolving traffic conditions, as they occur in the simulation.

The second aspect is associated with the representation of scale. Traditionally, traffic models have been designed to operate at relatively coarse spatial scales, such as the traffic analysis zone, and, arguably, the results that they generate are of relatively little value because of the scales at which they operate (Batty, 2001). For example, standard land use and transport models commonly ignore pedestrian traffic, in many instances because the available methodologies cannot adequately represent trips at that scale. However, the advances that we have already discussed have made the design of very detailed simulations possible, commonly at “microscopic” scales at the level of individual vehicles and pedestrians.

The third distinguishing factor, the ability to perform entity-based simulation (Benenson and Torrens, 2003), is closely related to this. The increase in resolution of traffic simulation has made it possible to abandon the idea of a “mean individual,” with average behaviors and characteristics derived from those of the group. Simulated entities can be designed, instead, at an “atomic” level (Anderson, 2002), with entities represented in terms of their distinct individual attributes and behavior (Gimblett, 2002). This has important implications for circumnavigating problems of ecological fallacy in model development, because the models can be run with spatially non-modifiable units.

The fourth characteristic is interaction and its representation. It is now possible to move beyond a reliance on interaction as a flow between modeled entities – an approach characterized by gravity and spatial interaction models – and into the representation of more localized interaction. Higher-level interactions can also be represented, often seamlessly, as they emerge from more micro-scale activity. In addition, the geosimulation approach allows model developers to abandon the notion that interactions take place evenly across a system (Anderson, 2002).

Fundamentally, this has resulted in a paradigm shift in traffic simulation. There is now a sense of using geosimulation-style environments as tools for hypothesis testing and “what if” scenario exploration, but with an unprecedented degree of realism.

#### 4. Automata as geosimulation tools

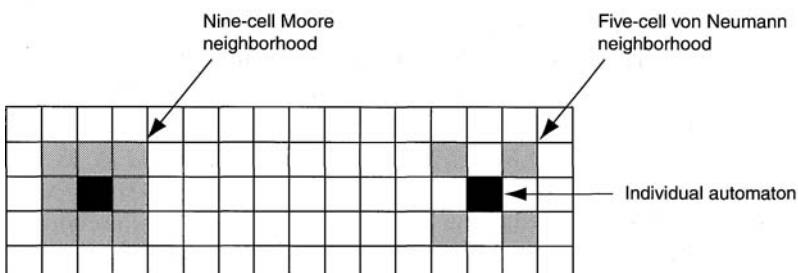
The geosimulation approach is perhaps best represented in automata-based modeling. Automata-based tools such as CA and MAS encapsulate all of the features of the geosimulation approach described in the previous sections. Both tools are used to develop geosimulation-style traffic models.

Automata are general processing units, most often artificial in design. They can be endowed with characteristics that change over time based on the internal attributes of the automaton itself, a set of transition rules, and input from outside the automaton. Automata provide a formal mechanism for expressing the fundamental elements of a system and the nature of their interactions. Mathematically, an automaton can be described with a few symbols:

$$S_{t+1} = f(S_t, I_t).$$

States ( $S$ ) describe the attributes of an automaton at a given point in time ( $t$ ). Transition rules, expressed here as a functional statement ( $f$ ), govern how the state of an automaton should change from time  $t$  to a subsequent period in time ( $t + 1$ ). The transition calculation is based on the state of the automaton itself at time  $t$ , as well as information gleaned from an input ( $I$ ) of some description, usually derived from the states of neighboring automata, introduced at time  $t$ .

CA and MAS are extensions of this basic idea. In CA, individual automata are interpreted as being housed within a cellular boundary, such as a grid square. Together, these “cells” form a continuous lattice of connected automata, and individual automata are fixed in this lattice. External input to particular automata is derived from a neighborhood of cells within a localized area of the lattice around an automaton (Figure 1). With MAS, individual automata are themselves free to move in space; they are mobile (Figure 2). Furthermore, the attributes that describe individual automata generally attribute some agent-like qualities to the unit, such as anthropomorphic characteristics, and the transition rules that govern



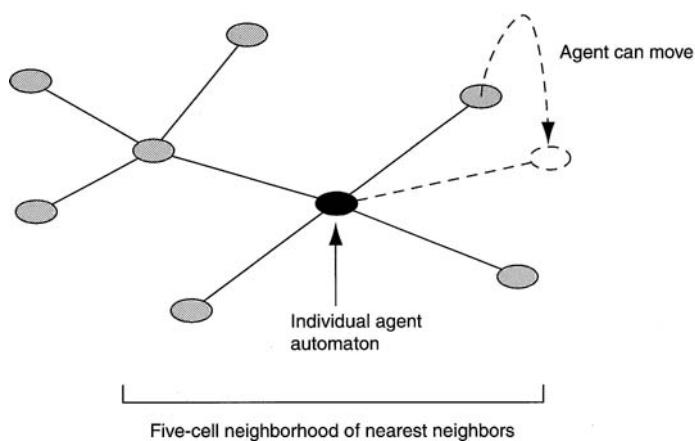


Figure 2. Multi-agent systems.

change in MAS are usually interpreted to represent agent-like behaviors, such as preference formulation, walking movement, driving behavior, etc.

Automata-based tools such as CA and MAS encapsulate all of the features of the geosimulation approach, and provide a methodology for representing them in a simulation framework. Time is handled through the transition function, which determines how automata states change dynamically. A variety of scales can be accommodated, since automata can be designed to represent units of any desired resolution. The entity approach is facilitated through the independent and discrete specification of individual automata. Finally, interaction is enabled through the neighborhood function, which determines how independent automata units should react to, and interact with, neighboring units.

Hopefully, it is easy to envisage how various traffic systems can be represented as interacting collections of automata. Automata can be designed to represent any unit of a traffic system: vehicles, pedestrians, sections of sidewalks, etc. These entities can be endowed with various attributes of relevance to traffic: velocity, demographic characteristics, speed limits, etc. Automata might be associated with various lattices, designed to represent features of a transport system: regular grid-based tessellations, irregular grids, graph-based networks of nodes and edges, etc. Also, neighborhood functions can be used to mimic features of traffic systems, such as gaps between vehicles, spaces for overtaking in adjacent lanes, pedestrians' perception of the space immediately surrounding them, etc. Transition rules can be designed to represent an almost limitless array of behaviors and processes: lane-changing rules, rules describing motion, collision avoidance, etc. In addition, a variety of time scales can be introduced into the models.

In the sections that follow, we will outline the use of geosimulation-style automata tools to model vehicular and pedestrian traffic systems. The various ways in which spatial topology, entity descriptions, neighborhood definitions, time, and transition rules are encoded into the models will be discussed.

## 5. Modeling vehicular traffic

The geosimulation approach allows for “microscopic” traffic modeling, with individual vehicles being simulated as independent entities, and permits for the simulation of interactions between those vehicles along simulated roads. Often, recognizable traffic conditions, such as congestion (Nagel and Schreckenberg, 1995), emerge from these interactions and in many cases the models provide mechanisms for simulated entities to react and adapt to these conditions as they evolve in near real time.

Automata-based traffic models are developed in much the same way as those described in the previous sections. The spatial or network structure of the traffic environments that are being simulated are encoded into the model as lattices. Simulated entities are designed with various characteristics that enable them to function in a manner similar to their real world counterparts. Neighborhoods of influence designate the spatial domain of interactions between entities in the simulation. Some form of internal clock is introduced into the simulation, allowing for dynamic action in the model. Conditional rules and calculations are also included in the simulation, describing how modeled entities should perceive their modeled environment, react to changes in their own state descriptions, react to other entities in the simulation, and react to changes in their environment. Automata, either CA or MAS, are the principal mechanism used for the simulation of vehicles and the environment in which they travel.

### 5.1. Spatial topology

Roads are encoded into automata models in familiar ways: lattice nodes represent road junctions and links represent roads that connect those junctions. Additional detailed topology may also be introduced. In the TRANSIMS model, for example, land use and connectivity data, intersections (signs and signals), activity locations, parking, transit stops, and route paths are also encoded into the topology of the model (Barrett et al., 1999).

Queues are used to represent the vehicles that travel along a link. Queues are commonly encoded as one-dimensional lattices (or parallel lattices where multi-lane roads are represented), with each cell in the lattice represented as a cellular automaton. Where models are developed for experimental purposes, such as

studying the formation of congestion in an abstract sense, queues may be coded with periodic boundary conditions, i.e. traffic moves in a continuous loop through the queue (Rickert et al., 1996). Various characteristics can be associated with the cells that form a queue, e.g. length, flow capacity, free flow velocity, and free flow travel time (Cetin et al., 2001). In this way, automata cells and lattices can be used to build realistic traffic environments.

### *5.2. Entity descriptions*

One of the great advantages of the geosimulation approach is that it allows simulated entities to be represented as atomic objects. Whereas spatial interaction models represent aggregate flows of traffic, geosimulation models represent the individual particles that comprise that flow: individual cars and trucks and their drivers for vehicular traffic, and individual walkers for pedestrian traffic. In most of the microscopic traffic models, vehicles are encoded as individual cells of 7.5 m in length, which is the length of a car plus the gap between cars in a jam (Wagner et al., 1997).

Simulated entities can be afforded a rich range of state descriptors describing their characteristics. There is no need for “mean individual” descriptions; entities can be represented as true individuals. In the PARAMICS model, for example, individual vehicles are encoded with state variables that represent a vehicle’s length, maximum acceleration and deceleration, cornering speed, desired destination, preferred traveling speed, current position, and current direction (Wylie et al., 1993). The TRANSIMS model is also capable of representing much of that information, but adds even more detail to the description of vehicles, including a record of the household to which the vehicle belongs, the initial network location of the vehicle, and a vehicle classification from a 23 type scheme (Barrett et al., 1999).

### *5.3. Neighborhood definitions*

Neighborhoods of influence can be defined for individual vehicle automata in the simulation. These neighborhoods represent the range of influence for interaction between modeled entities. Neighborhoods are used to model drivers’ “perception” of traffic conditions around them, such as the buffer between adjacent cars in the same lane, or gap opportunities for merging traffic at junctions (Wylie et al., 1993). Sophisticated neighborhood functions may also be introduced to facilitate lane-changing decisions, e.g. how far to look ahead or behind in the same lane and how far to look ahead in adjacent lanes before switching position (Rickert et al., 1996). Neighborhoods can be defined in static terms, e.g. occupancy of five cells ahead or

in front of a vehicle (Barrett et al., 1999). Or, alternatively, neighborhoods can be related to other dynamic characteristics of the model, such as the velocity of a moving vehicle (Rickert et al., 1996).

#### 5.4. Time

The ability to encode dynamic relations in a simulation model is another advantage of the geosimulation approach for traffic modeling, where users are often interested in moment-by-moment traffic dynamics for systems of interest. There are two ways in which the geosimulation approach is particularly innovative in relation to its treatment of time: temporal resolution and parallel update. Traffic applications of geosimulation-style modeling are among some of the most fine-scaled simulations, in terms of temporal resolution, in urban studies. This is partly because they need to be – the reaction time of drivers is of the order of 1 s (Wagner et al., 1997) – and is partly a function of the incredible computing power available to compile and run these models. Further advantages stem from the synchronous nature of transition rules in the models. In keeping with automata-based principles, traffic geosimulation models are often updated in parallel (on supercomputers, clustered processors, or networks of machines); transition rules are applied to modeled entities, and their states are altered in unison, throughout the simulation. When coupled with a fine-scale temporal resolution, this enables the simulated system to “evolve” dynamically, analogous to real-world conditions. In this sense, traffic geosimulation models can now be run, in many cases, in near real time for medium-sized cities. In addition, the reaction of individual vehicles to evolving traffic conditions (accidents, congestion, detours, etc.) can be simulated dynamically.

#### 5.5. Rules

Various transition rules can be used to characterize the behavior of vehicles, and their drivers, in automata-based traffic simulations. Of course, it would be a daunting task to formulate rules to mimic the full range of driving behaviors, so model developers focus on a minimal set of rules instead (Wagner et al., 1997). Just as in complexity studies, traffic simulations are designed with a few simple rules, and it is hoped that more complex behaviors will “emerge” through the myriad application of those rules between many interacting entities.

Traffic geosimulation models are noteworthy in their attention to rules of movement. Transition rules are formulated to simulate acceleration and braking as a function of various vehicle characteristics (speed, maximum velocity, target speed, etc.) and conditions in a vehicle’s neighborhood (type of road, perceived

traffic conditions ahead, gap to the next car, etc.) (Wylie et al., 1993; Wagner et al., 1997). In some instances, random acceleration and deceleration functions are also introduced to mimic erratic movement (Rickert et al., 1996). Rules for collision avoidance have also been introduced. Other rules have been developed to simulate signal stopping behavior (Barrett et al., 1999) and traffic movement at junctions, with “gap acceptance” functions that determine how long vehicles must wait at a junction before they can proceed (Wylie et al., 1993). In models where collections of vehicles are simulated as traffic queues (Rickert et al., 1996; Barrett et al., 1999), entrance and departure from vehicle queues can also be simulated, with vehicles leaving a queue freeing up space on a link, allowing another vehicle to join the queue.

Quite elaborate rules have also been devised to simulate lane-changing behavior. In this sense, automata models resemble traditional queuing lane models. However, traditional queuing lane models are not truly multi-lane in their design (Cetin et al., 2001); they approximate multiple lanes by switching the positional order of vehicles to make it appear as if passing has occurred. In automata models, however, parallel lattices can be constructed adjacent to each other, facilitating the simulation of movement between lanes. Lane-changing rules in automata traffic models can then simulate exchanges of vehicles between lanes as a function of a variety of factors, including the number of empty sites in a vehicle’s neighborhood ahead in the same lane, ahead in adjacent lanes, and behind in adjacent lanes; velocity; and hindrance in the current lane (Rickert et al., 1996; Wagner et al., 1997).

## 6. Modeling pedestrian traffic

Traditionally, pedestrian traffic has been comparatively ignored by transport modelers. There are many reasons why this may have been the case (Batty, 2001). To a certain degree, pedestrian traffic has been overshadowed by vehicular traffic as an area of applied research. The demand for vehicular transport, at least in contemporary metropolitan areas in developed countries, outpaces that for pedestrian modes of travel by a significant margin. Likewise, the multitude of problems – such as environmental, health, and social justice – associated with vehicular transport overshadows those tied to pedestrian travel. Scale issues also factor into the relative favor afforded vehicular transport. The range of movement permitted by vehicular transport, and the associated scale of its influence, are far greater than that of pedestrian movement. Vehicular transport problems are also, to some extent, more “tractable” than pedestrian transport problems, partly because of the aforementioned scaling issues, and partly because of the greater attention paid to vehicles and the wider array of modeling techniques that are available.

However, in recent years, the landscape for pedestrian transport research has improved considerably. This is partly a response to shifts in the political agenda in relation to transport, particularly heightened awareness of “sustainability” in urban environments and modes of transport. As in other area of transport modeling, the field has also benefitted from innovations in the research landscape. The development of geosimulation-style approaches, however, has perhaps been most significant in initiating the recent flurry of research in pedestrian modeling. This has been supported by the development of new data capture techniques for pedestrian models: aerial photography for capturing crowd volumes and movement through automated observation, time lapse filming, video data, and intelligent image-processing techniques for extracting information from these data.

Together, these advances have enabled the development of innovative, microscopic, “agent-based” pedestrian simulations in which the activity schedules and second-by-second movement and interactions of individual walkers are simulated, sometimes for large crowds of pedestrians in whole districts of a city. These models enable applied work to be performed that had either been previously intractable or not imagined at all.

Pedestrian traffic modeling is, in many respects, a far more complex simulation problem than vehicle traffic modeling. This is particularly true at “microscopic” levels. The scale of observation can often be much finer for pedestrian modeling, simply because the “footprint” of a pedestrian is generally much smaller than that of a vehicle. Furthermore, the behavior of pedestrians is not as constrained as that of vehicles on roads. There are generally many more paths available to pedestrians when compared to vehicles. Pedestrians are also much less limited in their range of movement; they can, for example, perform side-step and about-face maneuvers. Pedestrians are not generally constrained by rules of the road; they can, for example, ignore crossing rules at intersections by jaywalking. Finally, pedestrians themselves, as well as their behavior, are much more varied than vehicles, at least in a general sense. Despite the age, gender, social, cultural, and health characteristics of various drivers, most cars behave in a relatively similar manner on the road. That is not true of pedestrian walkers.

For these reasons, geosimulation-style techniques are ideally suited to modeling pedestrian traffic. MAS, in particular, are well-suited to the task. The comparative flexibility of MAS tools compared to CA, with respect to representing movement and interaction, makes them ideal for representing complex adaptive phenomena like pedestrian crowds.

### *6.1. Entities*

Generally, geosimulation-style pedestrian traffic models provide for the representation of two types of entities: agent pedestrians, and pedestrian obstacles in the built

environment. The simulated vehicle drivers discussed in the previous section could have various demographic and socio-economic state variables associated with them. This is also the case with pedestrian traffic models, where simulated agent pedestrians are often attributed various life-like characteristics to help shape their movement behavior and to populate the models with agents that are likely to behave in a diverse fashion (Haklay et al., 2001). Other characteristics of relevance to traffic modeling can be observed as agents move within the simulation, e.g. position, direction, time in the system, movement, and states, and this has close analogies to other pedestrian-flow-modeling approaches (Hoogendoorn and Bovy, 2002).

State variables can also be ascribed to various entities used to represent the physical environment in which pedestrian agents interact, both in terms of attraction features (buildings, shops, etc.) and potential obstacles (street furniture, walls, road signs, etc.) (Kerridge et al., 2001).

## 6.2. *Spatial topology*

Invariably, grid-based lattices are used to represent the spatial topology of the environments in which agents interact, as is the case in vehicle traffic models. Of course, a finer resolution of representation is often necessary for pedestrian models; in some instances grid squares have been used to represent spaces of  $750 \times 750$  mm in size. Various features of the built environment – building outlines, land uses, divisions between sidewalks and roads, network data, gateways and waypoints, etc. – may be embedded into this typology, either as raster or vector data. Also, various representations of street and building layouts can be altered in the model structure to allow for the evaluation of planning and design issues relating to pedestrian movement.

## 6.3. *Time*

As in most geosimulation-style models, time is generally discrete in pedestrian traffic simulations, proceeding in “chunks” that approach real time. However, discrete units of time are commonly designed at very fine temporal scales. In the PEDFLOW model (Kerridge et al., 2001), for example, time is divided into slots of 0.1 s in duration.

## 6.4. *Neighborhoods*

Various neighborhood functions may be introduced to determine pedestrian agents’ “awareness” of conditions in the environments surrounding them, both for

the detection of targets and potential obstacles and the determination of avenues for collision avoidance. In their models of agent-based shoppers Turner and Penn (2002) specify agents with neighborhoods derived from their lines of sight. In the STREETS model (Haklay et al., 2001), agents "look" in up to five directions in their immediate vicinity to determine where the most space is available for movement. In the PEDFLOW model (Kerridge et al., 2001), agents are equipped with three neighborhood filters: a "static awareness" function that determines how far ahead an agent can "see"; a "preferred gap size" that governs the smallest space a pedestrian is willing to move into; and a "personal space" function that sets the amount of buffering space a pedestrian would like to maintain around itself. These neighborhood functions provide simulated pedestrians with the spatial "cognition" necessary for realistic movement patterns.

### *6.5. Rules*

It has been noted that, as is the case with vehicular traffic movement, there is an almost limitless array of behaviors and factors that contribute to the movement dynamics of a pedestrian crowd. Nevertheless, pedestrian movement is surprisingly predictable. Despite the apparent chaos of crowd dynamics, certain regularities can be observed, and these can be used to formulate transition rules to drive movement behavior in agent-based simulations. Helbing et al. (2000, 2001) detail several of these regularities. Pedestrians usually pursue the fastest route to a target destination, and prefer to travel at the most comfortable walking speed. Pedestrians also like to maintain a buffering distance from other pedestrians and obstacles. Certain automatic behaviors may also be observed in certain situations, e.g. when entering congested doorways or side-stepping obstacles. Also, at a more macro-level, gas- or fluid-like qualities may be attributed to pedestrian crowds at certain densities, and similarities with granular flows have also been noted. These observations may be used to formulate transition rules governing the speed and movement of pedestrian agents in traffic simulations.

In the PEDFLOW model (Kerridge et al., 2001), for example, the speed of pedestrian agent movement is determined by factoring in the time period over which an agent occupies a grid square, proportional to its own walking speed or that of other pedestrian agents in its neighborhood. In the STREETS model (Haklay et al., 2001), pedestrian agents are endowed with maximum walking speed attributes and categorical variables that characterize their speed at any given moment (e.g. "stuck," "standing," and "moving"), and these variables are used to determine the speed at which a simulated pedestrian walks.

Agent movement – wayfinding and navigation – is determined by rules that are analogous to those for vehicular traffic. Activity models determine the overall movement schedules of agents and any associated target destination or waypoints.

(Agents may decide to adhere to those schedules or wander from pre-assigned targets.) Various navigation rules then determine how agents navigate to those destinations within their simulated environments, reacting to and interacting with the emerging dynamics within the simulated system. In the STREETS model (Haklay et al., 2001), for example, various “helmsman” rules are used to mediate between an agent’s “best heading” and its desired target, while “navigator” rules maintain agents’ overall heading in relation to targets. On a more micro-scale, rules are often introduced to determine how pedestrian agents should react to evolving conditions in their immediate surroundings: to determine step-by-step movement and collision detection. In the STREETS model, a “walkability” calculation is performed to assess whether enough space exists ahead of an agent for it to proceed along its route. Agents then move to grid squares with the most “walkability.” In the PEDFLOW model (Kerridge et al., 2001), agents perform similar calculations in relation to their neighborhoods, distinguishing between observed entities in that neighborhood (other pedestrian agents, goal points, stationary obstacles, buildings, and kerbs). Agents calculate the distance to those objects, and then execute one of four actions to proceed: go straight ahead, go diagonally left or right, move to the side (a choice parameter determines which side they favor), or remain where they are. Using these rules, pedestrian agents can be designed, choreographically, to mimic the movement patterns of real walkers, both at an individual scale and as a crowd.

## 7. Conclusion

We have just described how automata-based tools can be used, as geosimulation-style models, to simulate detailed and dynamic interactions in vehicular and pedestrian traffic systems. The limitations associated with these tools in urban simulation have been addressed elsewhere (Batty and Torrens, 2001; Torrens and O’Sullivan, 2001). While the introduction of these tools to traffic simulation has several advantages and offers much potential for the development of more realistic and useful traffic simulation environments, a number of hurdles to their widespread deployment remain.

Existing theory about traffic systems may be inadequate for developing geosimulation-style traffic models. In particular, further development may require new understandings of interactive traffic behavior, in particular of how it organizes from micro- to macro-scales. This is particularly true in relation to pedestrian modeling, which, as we have seen, has not been as actively pursued as vehicle traffic modeling. Nevertheless, significant new insights are being made, particularly in relation to observed regularities in pedestrian movement patterns (Helbing et al., 2000, 2001) and pedestrian choice heuristics (Kurose et al.,

2001). Geosimulation tools may well play a role in exploring and validating new theories.

Because of their fine resolution and dynamic nature, geosimulation models often require large amounts of computing resources to run. Despite recent developments in computer hardware, the volume of entities and interactions represented in geosimulation-style models makes them extremely "resource hungry" in terms of computing power. Processing power, in particular, is still weak in many instances. Advances in the processing power of traffic simulations is being made, particularly in relation to networking computers and harnessing their combined processing power (Nagel and Rickert, 2001), but research into this field is likely to continue.

Other issues remain regarding the application of geosimulation-style models in practice. To a certain extent, many of these simulations, particularly those developed for pedestrian traffic, are academic in nature. Despite popular applications to case studies (Barrett et al., 2001), the tools have yet to enjoy widespread testing in real-world contexts.

Data considerations are also important. In some instances, detailed data sets exist to "feed" geosimulation-style models, but generally fine-scale data are not readily available. This makes the validation and calibration of geosimulation-style models a difficult task. It is likely that new data sources will become available, but, in the meantime, research is uncovering innovative approaches to the dataware problem. The generation of data sets containing "synthetic" households and vehicles (Bush, 2001) is one such avenue of research inquiry.

Nevertheless, despite these caveats, the geosimulation approach offers promising avenues for traffic model development. The introduction of automata-based tools, in particular, to traffic simulation, has enabled the construction of a new class of simulation. These environments enable the incorporation of a diversity of theoretical ideas about traffic systems directly into the simulation framework, and facilitate the development of richly detailed and dynamic simulation environments. The field is in its early stages of exploration, but the indications for the development of innovative tools, testing of new theories, and application of simulation to applied contexts look promising.

## References

- Anderson, J. (2002) "Providing a broad spectrum of agents in spatially explicit simulation models: the Gensim approach," in: H.R. Gimblett, ed., *Integrating geographic information systems and agent-based modeling techniques for simulating social and ecological processes*. Oxford: Oxford University press.
- Barrett, C.L., R.J. Beckman, K.P. Berkbigler, K.R. Bisset, B.W. Bush, S. Eubank, J.M. Hurford, G. Konjevod, D.A. Kubicek, M.V. Marathe, J.D. Morgeson, M. Rickert, P.R. Romero, L.L. Smith, M.P. Speckman, P.L. Speckman, P.E. Stretz, G.L. Thayer and M.D. Williams (1999) *TRANSIMS*

- (transportation analysis simulation system), Vol. 0. Overview, Report LA-UR-99-1658. Los Alamos: Los Alamos National Laboratory.
- Barrett, C.L., R.J. Beckman, K.P. Berkbigler, K.R. Bisset, B.W. Bush, K. Campbell, S. Eubank, K.M. Henson, J.M. Hurford, D.A. Kubicek, M.V. Marathe, P.R. Romero, J.P. Smith, L.L. Smith, P.E. Stretz, G.L. Thayer, E. Van Eckhouwt and M.D. Williams (2001) *Transportation analysis simulation system (TRANSIMS)*, Reports LA-UR-01-5711, LA-UR-01-5712, LA-UR-01-5713, LA-UR-01-5714 and LA-UR-01-5715. Los Alamos: Los Alamos National Laboratory.
- Batty, M. (2001) "Agent-based pedestrian modeling," *Environment and Planning B*, 28:321–326.
- Batty, M. and P.M. Torrens (2001) "Modeling complexity: the limits to prediction," *CyberGeo*, 201.
- Benenson, I. and P.M. Torrens (2003) "Geosimulation: object-based modeling of urban phenomena," *Computers, Environment and Urban Systems*, Special issue.
- Bush, B.W. (2001) *Portland synthetic population*, Report LA-UR-00-5972. Los Alamos: Los Alamos National Laboratory.
- Cetin, N., K. Nagel, B. Raney and A. Voellmy (2001) "Large-scale multi-agent transportation systems," *Computational Physics Communications*, 147:559–564.
- Gimblett, H.R. (2002) "Integrating geographic information systems and agent-based technologies for modeling and simulating social and ecological phenomena," in: H.R. Gimblett, ed., *Integrating geographic information systems and agent-based modeling techniques for simulating social and ecological processes*. Oxford: Oxford University Press.
- Gurney, K. (1997) *An introduction to neural networks*. London: Routledge.
- Haklay, M., D. O'Sullivan, M. Thurstain-Goodwin and T. Schelhorn (2001) "So go downtown": simulating pedestrian movement in town centres," *Environment and Planning B*, 28:343–359.
- Helbing, D., F. Illés and T. Vicsek (2000) "Simulating dynamical features of escape panic," *Nature*, 407:487–490.
- Helbing, D., P. Molnár, I. Farkas and K. Bolay (2001) "Self-organizing pedestrian movement," *Environment and Planning B*, 28:361–383.
- Holland, J.H. (1998) *Emergence: from chaos to order*. Reading: Perseus Books.
- Hoogendoorn, S.P. and P.H.L. Bovy (2002) "Normative pedestrian behaviour theory and modelling," in: M.A.P. Taylor, ed., *Transportation and traffic theory in the 21st century*. Oxford: Pergamon.
- Kerridge, J., J. Hine and M. Wigan (2001) "Agent-based modelling of pedestrian movements: the questions that need to be asked and answered," *Environment and Planning B*, 28:327–341.
- Kurose, S., A.W.J. Borgers and H.J.P. Timmermans (2001) "Classifying pedestrian shopping behaviour according to implied heuristic choice rules," *Environment and Planning B*, 28:405–418.
- Kurzweil, R. (1990) *The age of intelligent machines*. Cambridge: MIT Press.
- Kurzweil, R. (1999) *The age of spiritual machines: how we will live, work and think in the new age of intelligent machines*. London: Phoenix.
- Levy, S. (1992) *Artificial life: the quest for a new creation*. London: Penguin Books.
- Miller, H.J. and S.-L. Shaw (2001) *Geographic information systems for transportation: principles and applications*. Oxford: Oxford University Press.
- Nagel, K. and M. Rickert (2001) "Parallel implementation of the TRANSIMS micro-simulation," *Parallel Computing*, 27:1611–1639.
- Nagel, K. and M. Schreckenberg (1995) *Traffic jams in stochastic cellular automata*, Report 95ATS089. Los Alamos: Los Alamos National Laboratory.
- Rickert, M., K. Nagel, M. Schreckenberg and A. Latour (1996) "Two lane traffic simulations using cellular automata," *Physica A*, 231:534–550.
- Schleifer, R. (2002) "Intelligent agents in traffic and transportation," *Transportation Research C*, 10:325–329.
- Sipper, M. (1997) *Evolution of parallel cellular machines: the cellular programming approach*. Berlin: Springer-Verlag.
- Thill, J.-C., ed. (2000) *Geographic information systems in transportation research*. Oxford: Pergamon.
- Torrens, P.M. (2002) "Cellular automata and multi-agent systems as planning support tools," in: S. Geertman and J. Stillwell, eds, *Planning support systems in practice*. London: Springer-Verlag.
- Torrens, P.M. and D. O'Sullivan (2001) "Cellular automata and urban simulation: where do we go from here?" *Environment and Planning B*, 28:163–168.
- Turner, A. and A. Penn (2002) "Encoding natural movement as an agent-based system: an investigation into human pedestrian behaviour in the built environment," *Environment and Planning B*, 29:473–490.

- Wagner, P., K. Nagel and D.E. Wolf (1997) "Realistic multi-lane traffic rules for cellular automata," *Physica A*, 234:687-698.
- Wylie, B., G. Cameron, W. Matthew and D. McArthur (1993) "PARAMICS: parallel microscopic traffic simulator," in: *Second European Connection Machine Users Meeting*, Paper. Paris: Meudon.

***Part 9***

**NETWORKS**

## DESIGN AND ANALYSIS OF TRANSPORT NETWORKS

HAI YANG and XIAONING ZHANG

*The Hong Kong University of Science and Technology, Kowloon*

### 1. Introduction

Traffic congestion in urban areas is an urgent and tangible problem impinging on our society and the lives of each of us as individuals. This problem will become increasingly serious as traffic demands rise, leading to travel delays, increased fuel consumption, severe air pollution, etc. Therefore, it is vitally important to carefully design and control transportation networks in order to provide an efficient and reliable level of service for users. Broadly speaking, network design problems refer to the optimal selection of various alternatives to improve the performance of the transportation system under certain budget constraints, including, as examples, road capacity expansion, road pricing, and traffic signal control. Financial investment on link capacity expansion is a direct, traditional way to mitigate traffic congestion; road pricing has long been recognized as an efficient means for travel demand management; and by optimizing signal settings, traffic engineers can provide smoother movement of traffic flow and enhance the reserve capacity of a signal-controlled road network.

Mathematically, the technique of network design and analysis is characterized by the problem of transportation network optimization with user equilibrium constraints (TNO-UEC), which has long been recognized to be one of the most difficult and challenging problems in the field of transportation science. Throughout the past two decades a vast and growing body of research has developed, focused on formulations for and solutions of the various TNO-UEC problems relating to efficient transportation planning and management. A common characteristic of these problems is the determination of a set of optimal values of the decision variables for optimizing various transportation system performance measures, while taking into account the route choice behavior of network users. Typical examples include the network expansion problem (Abdulaal and LeBlance, 1979; Yang and Bell, 2001), the road pricing problem (Yang and Lam, 1996), and the optimal signal control problem (Wong and Yang, 1997). This broad class of transportation problems can be characterized as a

Stackelberg game – sometimes called a leader and follower game – where the leader is a transportation planner making traffic planning and control decisions to optimize a system-wide performance function, and the followers are the network users who choose their routes in a user-optimal manner in response to the planner's decision. From the viewpoint of mathematical modeling, the bi-level programming approach is appropriate for formulating solutions to TNO-UEC problems in which the upper-level optimization problem is to minimize or maximize a system-wide objective function subject to certain constraints, while the lower-level problem is a non-linear programming model or variational inequalities (VI) formulation of the network user equilibrium problem. The problems are termed mathematical programs with equilibrium constraints (MPEC) in the literature when the equilibrium constraints are generally represented by VI or a non-linear complementarity problem. Due to the inherent non-convexity and non-differentiability of the reaction functions, the generalized non-linear bi-level programming problem or the MPEC problem is difficult to solve. In fact, the MPEC problem has become a major focus of operations researchers. Significant progress on the optimality conditions and algorithmic development of MPEC problems has been made, and is well documented.

In bi-level TNO-UEC problems, the lower-level user equilibrium problem can be formulated as optimization or VI problems in link flows, but path flows as intermediate variables have to be utilized. Since the number of feasible paths is huge for real transportation networks and the deterministic equilibrium path flow pattern is not unique, it is impracticable to describe the lower-level problem in path flows only when solving TNO-UEC problems. This implies that the most theoretical results developed in the mathematical literature cannot be used to solve TNO-UEC problems in an efficient and straightforward manner. This has motivated transportation scientists to design efficient solution methods by identifying the desirable properties of TNO-UEC problems, particularly with regard to the nature of the lower-level network equilibrium problem. For comprehensive reviews on the state-of-the-art developments of the models and algorithms for TNO-UEC problems, readers are referred to Migdalas (1995) and Yang and Bell (1998, 2001). Although various heuristic or theoretically sound algorithms are currently available (Yang and Bell, 2001; Patriksson and Rockafellar, 2002), there remains a great need for the development of efficient methods that can handle networks of realistic sizes.

This chapter is devoted to the design and analysis of transportation networks. In particular, various optimization models and efficient solution algorithms, together with simple illustrative examples, are presented. In Section 2, a general framework for the bi-level model in network design and analysis is provided, and alternative specific problems in network design are formulated. In Section 3, the mathematical properties of the bi-level model are described, and a

mathematically appealing approach, developed recently, is introduced to solve the bi-level model. Conclusions are provided in Section 4.

## 2. Formulations of network design problems

### 2.1. General framework of the bi-level model in network design and analysis

Let  $G \in (N, A)$  be a transportation network consisting of a set  $N$  of nodes and a set  $A$  of directed links. Each link  $a \in A$  has an associated flow-dependent travel cost function  $t_a(v_a)$ , where  $v_a$  is the traffic flow on link  $a \in A$ . We assume that the link cost function is a strictly increasing and continuously differentiable function in its own flow. Let  $W$  denote the set of origination–destination (O-D) pairs,  $R_w$  the set of all paths between O-D pair  $w \in W$ , and let  $f_r^w$  denote the traffic flow on path  $r \in R_w$ ,  $w \in W$ . The topology relationship of a link  $a \in A$  and a path  $r \in R_w$ ,  $w \in W$  is expressed by  $\delta_{ar}^w$ , where  $\delta_{ar}^w = 1$  if path  $r$  between O-D pair  $w \in W$  uses link  $a$ , and  $\delta_{ar}^w = 0$  otherwise.

The TNO-UEC problem is to choose a set of optimal values for decision variables to minimize (or maximize) an objective function while taking account of the route choice behavior of network users. The problem can be characterized as the general non-linear bi-level mathematical programming problem

$$\min_x F(v(x), x), \quad (1)$$

subject to

$$g_m(v(x), x) \leq 0, \quad m = 1, \dots, p_1, \quad (2)$$

$$h_n(v(x), x) \leq 0, \quad n = 1, \dots, p_2, \quad (3)$$

where  $x \in (\dots, x_k, \dots)^T$  is the vector of the upper-level decision variables, and  $v(x) \in (\dots, v_a(x), \dots)^T$  is the optimal solution of the following lower-level strictly convex optimization problem for any given  $x$ :

$$\min_v \sum_{a \in A} \int_0^{v_a} t_a(\omega, x) d\omega, \quad (4)$$

subject to

$$v_a = \sum_{w \in W} \sum_{r \in R_w} f_r^w \delta_{ar}^w, \quad a \in A, \quad (5)$$

$$\sum_{r \in R_w} f_r^w = q_w(x), \quad w \in W, \quad (6)$$

$$f_r^w \geq 0, \quad r \in R_w, w \in W, \quad (7)$$

where  $q_w(x)$  is the traffic demand between O-D pair  $w \in W$  that is dependent upon  $x$ ; functions  $g_m(v(x), x)$  and  $h_n(v(x), x)$  represent the constraints of the problem; and  $p_1$  and  $p_2$  denote the number of the inequality and equality constraints, respectively.

In the above bi-level programming framework (1)–(7), the upper-level decision variables enter and influence the lower-level problem either in the lower-level objective function (4) or the right-hand side of constraints (6). The objective function of the upper-level problem (1) is in general a performance function of the transportation network, and eqs (2)–(3) represent the constraints on the upper-level decision variables. The lower-level problem is an optimization formulation of the standard deterministic user equilibrium problem (Sheffi, 1985; Patriksson, 1994).

## 2.2. The continuous network design problem (CNDP)

For the CNDP, the decision variables are the capacity enhancements denoted by  $x_a$ ,  $a \in \bar{A}$ , where  $\bar{A} \subseteq A$  is the set of candidate links for capacity enhancement, the objective function is the total travel cost, and the constraint is a budget constraint. In this problem, the link travel cost is of course influenced by the corresponding capacity enhancement, and thus the link cost function is represented by  $t_a(v_a, x_a)$ ,  $a \in \bar{A}$ . Therefore, the bi-level programming model for the CNDP is specified as (Abdulaal and LeBlance, 1979)

$$\min_v \sum_{a \in A \setminus \bar{A}} t_a(v_a(x)) v_a(x) + \sum_{a \in \bar{A}} t_a(v_a(x), x_a) v_a(x), \quad (8)$$

subject to

$$\sum_{a \in \bar{A}} b_a(x_a) \leq B, \quad (9)$$

$$l_a \leq y_a \leq u_a, \quad a \in \bar{A}, \quad (10)$$

where  $\{v_a(x), a \in A\}$  is the solution of the following user equilibrium problem for given  $x \in (\dots, x_a, \dots)^T$ ,  $a \in \bar{A}$ :

$$\min_v \sum_{a \in A \setminus \bar{A}} \int_0^{v_a} t_a(\omega) d\omega + \sum_{a \in \bar{A}} \int_0^{v_a} t_a(\omega, x_a) d\omega, \quad (11)$$

subject to

$$v_a = \sum_{w \in W} \sum_{r \in R_w} f_r^w \delta_{ar}^w, \quad a \in \bar{A}, \quad (12)$$

$$\sum_{r \in R_w} f_r^w = q_w, \quad w \in W, \quad (13)$$

$$f_r^w \geq 0, \quad r \in R_w, w \in W, \quad (14)$$

where  $b_a(x_a)$  is an expanding cost (or investment cost) function for link  $a \in \bar{A}$ ,  $l_a$  and  $u_a$  are the lower and upper bounds of the allowable capacity enhancement for link  $a \in \bar{A}$ , and constant  $B$  is the total budget available for capacity enhancement.

### 2.3. The optimal toll pricing problem (OTPP)

For the OTPP, the decision variables are the tolls to be levied on a subset of links  $\bar{A} \subseteq A$ , and fewer objective functions can be adopted for the OTPP (Yang and Lam, 1996). Congestion pricing and private road tolling are two kinds of typical OTPPs. Total travel time is minimized in congestion pricing, or total revenue is maximized in private road tolling. Let  $\mathbf{x} \in (\dots, x_a, \dots)^T$ ,  $a \in \bar{A}$  denote the vector of link tolls and  $v(\mathbf{x}) \in (\dots, v_a(\mathbf{x}), \dots)^T$ ,  $a \in \bar{A}$  is the vector of the equilibrium link flows associated with link toll pattern  $\mathbf{x}$ . Then, the OTPP is formulated as

$$\max_{\mathbf{x}} \sum_{a \in \bar{A}} x_a v_a(\mathbf{x}) \quad (15)$$

or

$$\min_{\mathbf{x}} \sum_{a \in \bar{A}} t_a(v_a(\mathbf{x})) v_a(\mathbf{x}), \quad (16)$$

subject to

$$l_a \leq x_a \leq u_a, \quad a \in \bar{A}, \quad (17)$$

where  $v(\mathbf{x})$  is the solution of the following user equilibrium problem:

$$\min_{v} \sum_{a \in A} \int_0^{v_a} t_a(\omega) d\omega + \sum_{a \in \bar{A}} x_a v_a, \quad (18)$$

subject to constraints (12)–(14), where  $l_a$  and  $u_a$  are the lower and upper bounds of the toll levied on link  $a \in \bar{A}$ , respectively.

### 2.4. The signal-setting problem

The concept of reserve capacity has been extended from an isolated intersection to a general signal-controlled road network under time-stationary conditions. Wong and Yang (1997) proposed a bi-level programming method to determine signal setting for maximization of the network reserve capacity.

The set of signal-controlled intersections in a network is denoted by  $I$  ( $I \subset N$ ), and  $A_i$  denotes the set of links entering the signalized intersection  $i \in I$ , and  $\bar{A}$  denotes the set of all signal-controlled links  $\bar{A} = \{A_i, i \in I\}$ . Let  $\mathbf{q} \in (\dots, q_w, \dots)$  be the existing O-D matrix. Suppose that the existing O-D matrix is multiplied by a factor  $\mu$ , and thus the multiplied O-D matrix is  $\mu\mathbf{q}$ . Let  $x_i$  be a vector of timing variables associated with signalized intersection  $i \in I$ , and  $\mathbf{x} \in (\dots, x_i, \dots)$ . For a fixed and given  $\mathbf{q}$ , link flow  $v$  is a function of the demand multiplier  $\mu$  and signal-timing plan  $\mathbf{x}$ . The queues and delays at the intersections in the network under equilibrium conditions will be acceptable provided that the resulting degree of saturation on any link (signal-controlled links only are considered here, for simplicity) does not exceed a prescribed maximum acceptable value for that link. Namely,

$$v_a(\mu, \mathbf{x}) \leq p_a C_a(\mathbf{x}), \quad a \in \bar{A}, \quad (19)$$

where  $C_a(\mathbf{x})$  is the capacity of link  $X$ , which depends on the signal timings  $\mathbf{x}$ , and  $v_a(\mu, \mathbf{x})$  is the equilibrium flow for link  $a \in A$  that depends on the O-D demands and traffic signal settings, and  $p_a$  is the maximum acceptable degree of saturation for link  $a \in A$ . This constraint should be fulfilled, especially in networks with closely spaced intersections where queue spillbacks can be destructive, and avoidance of queuing is often an objective.

The signal-timing variables for links approaching a given signalized intersection  $i \in I$  should satisfy linear constraints that include cycle time, clearance time, minimum and maximum green times, etc., excluding the capacity constraint given by eq.(19). These constraints can be described in the following form:

$$\mathbf{G}_i x_i \geq \mathbf{b}_i \quad i \in I, \quad (20)$$

where matrix  $\mathbf{G}_i$  and vector  $\mathbf{b}_i$  depend on the specific timing specification for intersection  $i \in I$ , whether it is stage based or group based. For a detailed description, readers are referred to Allsop (1989).

Based on the aforementioned considerations and all the constraints, the problem to find the maximum value of O-D matrix multiplier can be formulated as the following bi-level programming problem:

$$\max_{\mu, \mathbf{x}} \mu, \quad (21)$$

subject to

$$v_a(\mu, \mathbf{x}) \leq p_a C_a(\mathbf{x}), \quad a \in \bar{A}, \quad (22)$$

$$\mathbf{G}_i x_i \geq \mathbf{b}_i, \quad i \in I, \quad (23)$$

where the equilibrium flow  $v_a(\mu, \mathbf{x})$ ,  $a \in A$  is obtained by solving the following network equilibrium problem:

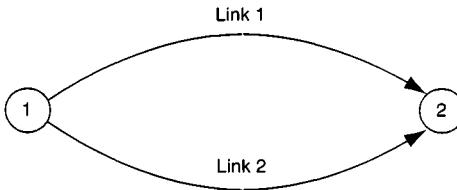


Figure 1. A simple network.

$$\min_v \sum_{a \in A} \int_0^{v_a} t_a(\omega, x) d\omega, \quad (24)$$

subject to

$$\sum_{r \in R_w} f_r^w = \mu q_w, \quad w \in W, \quad (25)$$

$$v_a = \sum_{w \in W} \sum_{r \in R_w} f_r^w \delta_{ar}^w, \quad a \in A, \quad (26)$$

$$f_r^w \geq 0, \quad r \in R_w, w \in W. \quad (27)$$

### 3. Properties of the bi-level model and the solution algorithm

#### 3.1. Non-differentiability of the reaction function

It is well known that bi-level models of the transportation network design problems are generally non-differentiable. Meng et al. (2001) provided a small example showing the non-differentiability of link flows as functions of link capacity expansion (termed the reaction or response function). Later, Meng and Yang (2002) further illustrated the non-differentiability of link flows as functions of O-D demand, in the problem of O-D estimation. To show the non-differentiability of the reaction function, here we provide a toll-pricing example for maximization of revenue. Consider a simple network depicted in Figure 1, consisting of two parallel links and two nodes with origin node 1 and destination node 2. Link 1 is an expressway allowing fast speeds and short free-flow travel time, and link 2 is a general highway with inferior performance. The expressway is subject to a toll charge, and the highway is free. The link cost functions for the two links are given by

$$t_1(v_1) \in 2v_1 + 2,$$

$$t_2(v_2) \in 4v_2 + 8.$$

Let the O-D demand from origin 1 to destination 2 be  $q_{12} \in 3$ . The road authority wants to maximize the income by adjusting the toll price  $x_1$  levied on link 1. The problem can be formulated as

$$\max_{x_1} x_1 v_1(x_1),$$

subject to

$$x_1 \geq 0,$$

where  $v_1(x_1)$  and  $v_2(x_2)$  are the optimal solution of the following lower-level strictly convex optimization problem for any given  $x_1$ :

$$\min_{(v_1, v_2)} \int_0^{v_1} (2\omega + 2) d\omega + \int_0^{v_2} (4\omega + 8) d\omega + x_1 v_1,$$

subject to

$$v_1 + v_2 \leq 3,$$

$$v_1 \geq 0, \quad v_2 \geq 0.$$

It is straightforward to find the equilibrium link flows  $v_1^*(x_1)$ ,  $v_2^*(x_1)$ , as follows:

$$v_1^*(x_1) = \begin{cases} 3 - x_1/6, & 0 \leq x_1 \leq 18, \\ 0, & x_1 > 18, \end{cases}$$

$$v_2^*(x_1) = \begin{cases} x_1/6, & 0 \leq x_1 \leq 18, \\ 3, & x_1 > 18. \end{cases}$$

Figure 2 plots the two equilibrium link flows as functions of the toll charge levied on link 1. Evidently, the two link flow functions are continuous in variable  $x_1$  but not differentiable at  $x_1 \in 18$ .

The above example illustrates that the TNO-UEC problem (1)–(7) in general is a non-differentiable optimization problem. This reveals the fact that most existing algorithms for non-linear programming problems are not directly applicable to TNO-UEC problems.

### 3.2. The marginal-function-based solution algorithm

We assume that  $q_w(\mathbf{x})$ ,  $w \in W$  are non-negative, continuously differentiable functions in  $\mathbf{x}$  and are bounded from above for a neighborhood of any given  $\mathbf{x}$ . Note that problem (4)–(7) is a special case of the general non-linear programming problem with perturbation parameter  $\mathbf{x}$ . The corresponding marginal function for this problem is defined by

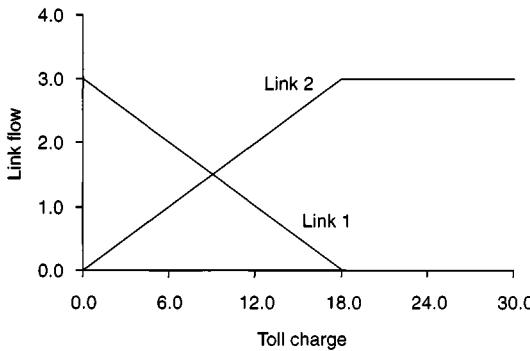


Figure 2. The equilibrium link flows as functions of toll charge on link 1.

$$\varphi(\mathbf{x}) = \min_{\mathbf{v} \in \Omega(\mathbf{x})} \phi(\mathbf{v}, \mathbf{x}), \quad (28)$$

where  $\Omega(\mathbf{x})$  is the set of feasible link flows associated with  $\mathbf{x}$ , satisfying constraints (5)–(7), and function  $\phi(\mathbf{v}, \mathbf{x})$  is defined as

$$\phi(\mathbf{v}, \mathbf{x}) = \sum_{a \in A} \int_0^{v_a^*(\mathbf{x})} t_a(\omega, \mathbf{x}) d\omega. \quad (29)$$

The marginal function (28) is a continuously differentiable function, and its gradient  $\nabla \varphi(\mathbf{x}) \in (\dots, \partial \varphi(\mathbf{x}) / \partial x_k, \dots)$  is given by

$$\frac{\partial \varphi(\mathbf{x})}{\partial x_k} = \sum_{a \in A} \int_0^{v_a^*(\mathbf{x})} \left( \frac{\partial t_a(\omega, \mathbf{x})}{\partial x_k} \right) d\omega - \sum_{w \in W} \pi_w^*(\mathbf{x}) \frac{\partial q_w(\mathbf{x})}{\partial x_k}, \quad k = 1, 2, \dots, \quad (30)$$

where  $\{v_a^*(\mathbf{x}), a \in A\}$  and  $\{\pi_w^*(\mathbf{x}), w \in W\}$  are the unique equilibrium link flow pattern and equilibrium O-D travel costs for a given  $\mathbf{x}$ , respectively.

Now we introduce the following gap function to characterize the user equilibrium conditions:

$$H(\mathbf{v}, \mathbf{x}) \in \phi(\mathbf{v}, \mathbf{x}) - \phi(\mathbf{x}), \quad \mathbf{x} \in X, \mathbf{v} \in \Omega(\mathbf{x}). \quad (31)$$

$X$  represents the feasible set of upper-level decision variables  $\mathbf{x}$ , and the function  $\phi(\mathbf{v}, \mathbf{x})$  is also defined in eq. (29). It is always true that the gap function  $H(\mathbf{v}, \mathbf{x}) \geq 0$ , and  $H(\mathbf{v}, \mathbf{x}) = 0$  if and only if the link flow pattern  $\mathbf{v}$  satisfies the user equilibrium conditions for a given  $\mathbf{x}$ . Furthermore,  $H(\mathbf{v}, \mathbf{x})$  is a continuously differentiable function, and its partial derivatives are given by

$$\frac{\partial H(\mathbf{v}, \mathbf{x})}{\partial v_a} = t_a(v_a, \mathbf{x}), \quad a \in A, \quad (32)$$

$$\frac{\partial H(\nu, \mathbf{x})}{\partial \nu_a} = \sum_{a \in A} \int_0^{v_a} \left( \frac{\partial t_a(\omega, \mathbf{x})}{\partial x_k} \right) d\omega - \frac{\partial \varphi(\mathbf{x})}{\partial x_k}, \quad k = 1, 2, \dots. \quad (33)$$

From the above properties, we can see that the equilibrium traffic assignment problem is equivalent to seeking a feasible link flow pattern  $\nu$  such that  $H(\nu, \mathbf{x}) \in 0$  for any given parameter  $\mathbf{x}$ . Therefore, the general bi-level programming model (1)–(7) is equivalent to the following conventional non-linear programming problem:

$$\min_{\nu, \mathbf{x}} F(\nu, \mathbf{x}), \quad (34)$$

subject to

$$g_m(\nu, \mathbf{x}) \leq 0, \quad m = 1 \dots, p_1, \quad (35)$$

$$h_n(\nu, \mathbf{x}) \leq 0, \quad n = 1 \dots, p_2, \quad (36)$$

$$H(\nu, \mathbf{x}) \in 0, \quad (37)$$

$$\nu_a = \sum_{w \in W} \sum_{r \in R_w} f_r^w \delta_{ar}, \quad a \in A, \quad (38)$$

$$\sum_{r \in R_w} f_r^w = q_w(\mathbf{x}), \quad w \in W, \quad (39)$$

$$f_r^w \geq 0, \quad r \in R_w, w \in W. \quad (40)$$

Note that the decision variables for the optimization problem (34)–(40) now include both link flows  $\nu$  and the original upper-level decision variables  $\mathbf{x}$ . This is contrasted with the original formulation (1)–(3), where  $\nu(\mathbf{x})$  is regarded as an implicit (reaction) function of  $\mathbf{x}$ , which is defined by the user equilibrium problem. Since the gap function  $H(\nu, \mathbf{x})$  is continuously differentiable, the optimization problem (34)–(40) is a typical single-level continuously differentiable non-linear programming problem if we assume that all functions  $F(\nu, \mathbf{x})$ ,  $g_m(\nu, \mathbf{x})$ ,  $m = 1 \dots, p_1$ , and  $h_n(\nu, \mathbf{x})$ ,  $n = 1 \dots, p_2$ , are continuously differentiable in  $(\nu, \mathbf{x})$ . From a computational viewpoint, the user equilibrium traffic assignment is sufficient to evaluate the value and gradient of function  $H(\nu, \mathbf{x})$  for a given feasible link flow pattern  $\nu$  and original decision variables  $\mathbf{x}$  according to eqs (28)–(33). These desirable properties allow the use of a large number of existing efficient non-linear programming algorithms and software packages to solve the reformulated, unified single-level TNO-UEC problem (34)–(40).

Although the non-differentiability issue of the original bi-level TNO-UEC problems has been resolved using the marginal function approach, the optimization problem (34)–(40) remains a non-convex optimization problem. In the following we consider how to generally find a local optimal solution or a

Karush–Kuhn–Tucker (KKT) stationary point of the problem by using existing efficient algorithms.

The functions  $g_m(v, x)$  and  $h_n(v, x)$  in constraints (35) and (36) take a linear form in real problems. Therefore, all the constraints except eq. (37) are linear in the problem. It is a natural way to impose the penalizing constraint (37) on to the objective function, and solve the new problem with non-linear penalized objective function and linear constraints. Using the above example, we will show how the method works.

In the above road-pricing example, after realizing the equilibrium link flows as functions of toll charge  $x_1$ , the revenue maximization problem can be rewritten as follows:

$$\max_{x_1 \geq 0} F(x_1) = \begin{cases} (3 - x_1/6)x_1, & 0 \leq x_1 \leq 18, \\ 0, & x_1 > 18. \end{cases}$$

This problem can be easily solved with an optimal solution  $X$  and  $F^* \in 13.5$ . Here, we remove the constraint  $v_1 + v_2 \leq 3$  by replacing  $v_2$  with  $3 - v_1$ . Now, consider its partially penalized problem.

The gap function defined above is given by

$$\varphi(x_1) = \min_{0 \leq v_1 \leq 3} \phi(v_1, x_1),$$

subject to

$$x_1 \geq 0.$$

Here,

$$\phi(v_1, x_1) = \int_0^{v_1} (2\omega + 2) d\omega + \int_0^{3-v_1} (4\omega + 8) d\omega + x_1 v_1 = 3v_1^2 - 18v_1 + x_1 v_1 + 42.$$

Solving the problem, the gap function is regarded as a function of toll charge, as below:

$$\varphi(x_1) = \begin{cases} -x_1^2/12 + 3x_1 + 15, & 0 \leq x_1 \leq 18, \\ 42, & x_1 > 18. \end{cases}$$

Evidently, this gap function is differentiable and its gradient is given by

$$\varphi'(x_1) = \begin{cases} -x_1^2/6 + 3, & 0 \leq x_1 \leq 18, \\ 0, & x_1 > 18. \end{cases}$$

This gradient is a continuous function of  $x_1$ , and thus the gap function is continuously differentiable. The variations of both the gap function and its gradient against toll charge  $x_1$  are plotted in Figure 3.

The partially penalized problem then becomes

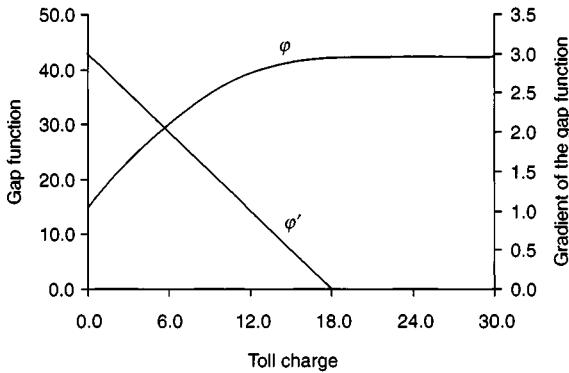


Figure 3. The variations of the gap function  $\varphi$  and its gradient  $\varphi'$  with toll charge  $x_1$ .

$$\max_{v_1, x_1} L(v_1, x_1, \rho) = F(v_1, x_1) + \rho(\varphi(x_1) - \phi(v_1, x_1)),$$

subject to

$$0 \leq v_1 \leq 3, \quad x_1 \geq 0,$$

where  $F(v_1, x_1) \in x_1 v_1$ , and  $\rho > 0$  is a penalty parameter. In view of the fact that the current objective function has different explicit expressions in different intervals of variable  $x_1$ , the penalty function can be decomposed into two subproblems defined within the intervals of  $0 \leq x_1 \leq 18$  and  $x_1 > 18$ , respectively.

When  $0 \leq x_1 \leq 18$ , the partially penalized objective function is quadratic, and the problem can be written as

$$\max_{v_1, x_1} A_1 \begin{pmatrix} v_1 \\ x_1 \end{pmatrix} + \frac{1}{2}(v_1, x_1) B_1 \begin{pmatrix} v_1 \\ x_1 \end{pmatrix} + C_1,$$

subject to

$$0 \leq v_1 \leq 3, \quad 0 \leq x_1 \leq 18,$$

where

$$A_1 = [18\rho, 3\rho], \quad B_1 = \begin{bmatrix} -6\rho & 1-\rho \\ 1-\rho & -\rho/6 \end{bmatrix}, \quad C_1 = -27\rho.$$

In this case,  $B_1$  is strongly concave for any  $\rho > 1.0$ , and the stationary point of the partially penalized problem of the example is

$$v_1(\rho) = \frac{3\rho}{2\rho-1}, \quad x_1(\rho) = \frac{18\rho}{2\rho-1},$$

which is feasible for the problem. Hence,  $(v_1(\rho), x_1(\rho))$  is the optimal solution, and we have

$$\lim_{\rho \rightarrow \infty} v_1(\rho) = 1.5, \quad \lim_{\rho \rightarrow \infty} x_1(\rho) = 9,$$

Note that for any limited value of  $\rho > 1.0$  the solution of the partial penalty problem will not be exactly equal to the solution of the example. Instead, the limit of the solution to the partially penalized problem is the solution of the example.

When  $x_1 > 18$ , the problem can be written as

$$\max_{v_1, x_1} A_2 \begin{pmatrix} v_1 \\ x_1 \end{pmatrix} + \frac{1}{2}(v_1, x_1)B_2 \begin{pmatrix} v_1 \\ x_1 \end{pmatrix} + C_2,$$

subject to

$$0 \leq v_1 \leq 3, \quad x_1 > 18,$$

where

$$A_2 = [18\rho, 0], \quad B_2 = \begin{bmatrix} -6\rho & 1-\rho \\ 1-\rho & 0 \end{bmatrix}, \quad C_2 = 0.$$

The optimal solution of the problem is not obvious here for a limited positive parameter  $\rho$ . Nevertheless, it is straightforward to check the optimal solution as  $\rho \rightarrow \infty$ . The problem can be rewritten as

$$\max_{v_1, x_1} x_1 v_1 - \rho v_1 (3v_1 + x_1 - 18),$$

subject to

$$0 \leq v_1 \leq 3, \quad x_1 > 18.$$

Clearly,  $v_1(3v_1 + x_1 - 18) \geq 0$  for any feasible  $v_1$  and  $x_1$ , and thus  $v_1(3v_1 + x_1 - 18) \in 0$  as the penalty parameter  $\rho \rightarrow \infty$ . This means that, as  $\rho \rightarrow \infty$ , the optimal value of  $v_1$  equals 0 but  $x_1 (> 18)$  can be any limited positive value greater than 18. The solution can be interpreted as follows. As long as the toll charge  $x_1$  on link 1 is greater than 18, no-one will use link 1. So the revenue is always 0, no matter how large the toll charge  $x_1 (> 18)$  is. This observation coincides with the result of the reaction function in Section 3.1.

Clearly, the optimal solution of the whole problem is located in the first interval and given by  $x_1^* = 9$ , and the corresponding maximal revenue is  $F^* \in 13.5$ . This basic example tells us that TNO-UEC problems can be solved using the marginal-function-based approach. A solution sufficiently close to the exact optimum can be sought by partially penalizing the continuously differentiable gap function into the objective function.

#### **4. Applications in location choice, land use, and network capacity**

After introducing the TNO-UEC modeling framework and algorithmic approach, we now describe two additional applications to the combined location and travel choice problem and the zonal and network reserve capacity problem.

It is well known that there is a trade-off between the travel cost of a journey and the work and residential benefits; thus, the design of a transportation network has to be jointly considered with location choice and land use issues. Recent combined location and travel choice models take into account the effect of traffic congestion by integrating urban location and travel choices in a unified, consistent equilibrium framework. Yang and Meng (1998) developed a mixed, combined, and stochastic user equilibrium model for urban travel and location choice problems with variable origin and destination location costs. The travel and location choices were given in logit models, which are based on random utility theory. A convex programming formulation of the problem was presented and a globally convergent algorithm for its solution was established. Meng et al. (2001) proposed a new combined transportation network equilibrium (for the demand side) and Lowry-type (Lowry, 1964) land use (for the supply side) model. The model takes into account the cost of the round trip to and from work in determination of trip distributions. The Lowry-type land use model is then used to determine the location of homes and jobs according to the round trip travel cost. The bi-level programming approach is used to determine the maximum number of trips that can be accommodated by the road network, subject to network capacity constraints.

Capacity modeling of a transportation network is an essential part of urban and transport systems planning. The capacity of a transport network indicates the maximum attainable throughput of the given network, and hence provides important information for efficient flow control and demand management. A good network capacity model would enable us to predict how much additional demand a road network could accommodate, and hence establish an efficient policy for traffic restraint and growth. Yang et al. (2000) investigated the modeling of road network capacity and level of service. A bi-level optimization formulation integrated with a combined trip distribution/assignment model was proposed to efficiently determine the maximum trip generation from each origin, subject to equilibrium network capacity constraints. The proposed model is contrasted with the previous methods by taking full account of the route and location choice behavior of travelers in determining the zonal and network reserve capacity.

#### **5. Conclusions**

This chapter presents a unified description of transportation network optimization problems with user equilibrium constraints. Specifically, the network link capacity

expansion problem, the road toll pricing problem and the optimal signal-timing problems are discussed. The traditional bi-level model formulation of these problems is transformed into a single-level continuously differentiable problem using a marginal function approach. An efficient partially penalized method is used to solve the resulting single-level differentiable optimization problems. The methods are illustrated with simple analytical examples. The marginal function approach proves to be very promising for dealing with transportation network optimization problems with user equilibrium constraints.

## References

- Abdulaal, M.S. and L.J. Leblance (1979) "Continuous equilibrium network design models," *Transportation Research* 13B, 19–32.
- Allsop, R.E. (1989) "Evolving applications of mathematical optimization in design and operation of individual signal-controlled road junctions," in: J.D. Griffiths, ed., *Mathematics in transport planning and control*. Oxford: Clarendon Press.
- Lowry, I.S. (1964) *Model of Metropolis*. Santa Monica: Rand.
- Meng, Q. and H. Yang (2002) "A unified continuously differentiable approach for the transportation network optimization problems with user equilibrium constraints," Working paper. Hong Kong: Hong Kong University of Science and Technology.
- Meng, Q., H. Yang and M.G.H. Bell (2001) "An equivalent continuously differentiable model and a locally convergent algorithm for the continuous network design problem," *Transportation Research B*, 35:83–105.
- Migdalas, A. (1995) "Bilevel programming in traffic planning: models, methods and challenge," *Journal of Global Optimization*, 7:361–405.
- Patriksson, M. (1994) *The traffic assignment problems: models and methods*. Utrecht: VSP.
- Patriksson, M. and R.T. Rockafellar (2002) "A mathematical model and descent algorithm for bilevel traffic management," *Transportation Science*, 36:271–291.
- Sheffi, Y. (1985) *Urban transportation networks: equilibrium analysis with mathematical programming methods*. England Cliffs: Prentice Hall.
- Wong, S.C. and H. Yang (1997) "Reserve capacity of a signal-controlled road network," *Transportation Research B*, 31:397–402.
- Yang, H. and M.G.H. Bell (1998) "Models and algorithms for road network design: a review and some new developments," *Transport Review*, 18:257–278.
- Yang, H. and M.G.H. Bell (2001) "Transport bilevel programming problems: recent methodological advances," *Transportation Research B*, 35:1–4.
- Yang, H. and W.H.K. Lam (1996) "Optimal road tolls under conditions of queuing and congestion," *Transportation Research A*, 32:319–332.
- Yang, H. and Q. Meng (1998) "An integrated network equilibrium model of urban location and travel choices," *Journal of Regional Science*, 38:575–598.
- Yang, H., M.G.H. Bell and Q. Meng (2000) "Modeling the capacity and level of service of urban transportation networks," *Transportation Research B*, 34:255–275.

*Chapter 33*

## SPATIAL EQUILIBRATION IN TRANSPORT NETWORKS

ANNA NAGURNEY

*University of Massachusetts, Amherst, MA*

### 1. Introduction

Transport networks are complex, large-scale spatial systems, and come in a variety of forms, ranging from road networks to air, rail, and waterway networks. They provide the foundation for the functioning of our economies and societies through the movement of people, goods, and services, and allow for the connectivity of residential locations with places of employment, schools, leisure activities, and retail outlets. From an economic perspective, the supply in such network systems is represented by the underlying network topology and the cost characteristics whereas the demand is represented by the users of the transportation system. An equilibrium occurs when the number of trips between an origin (e.g. residence/place of employment) and destination (place of employment/residence) equals the travel demand given by the market price, typically represented by the travel time for the trips.

The study of transport networks and their efficient management dates to ancient times. For example, the Romans imposed controls over chariot traffic at different times of day in order to deal with congestion (Banister and Button, 1993). From an economic perspective, some of the earliest contributions to the subject date to Pigou (1920), who considered a two-node, two-link transportation network, identified congestion as a problem, and recognized that distinct behavioral concepts regarding route selection may prevail (Knight, 1924).

The formal study of transport networks has challenged transportation scientists, economists, operations researchers, and engineers for several reasons: the above-mentioned size and scope of the systems involved; the behavior of the users of the network, which may vary according to the application setting, thereby leading to different optimality/equilibrium concepts; distinct classes of users may perceive the cost of utilizing the network in an individual fashion; congestion is playing an increasing role in numerous transport networks; and there may be interactions between transport and other foundational networks, such as telecommunications networks.

For example, to help one fix the size and scope of modern-day transport networks, the topology of the Chicago Regional Transportation Network consists of 12 982 nodes, 39 018 links, and 2 297 945 origin–destination (O-D) pairs of nodes between which travelers choose their routes (Bar-Gera, 1999), whereas in the Southern California Association of Governments' model there are 25 428 nodes, 99 240 links, 3217 O-D pairs, and six distinct classes of users (Wu et al., 2000).

Road congestion results in US \$100 billion in lost productivity in the USA alone, with the figure being approximately US \$150 billion in Europe, with the number of cars expected to increase by 50% by 2010 and to double by 2030 (Nagurney, 2000). Moreover, in many of today's transport networks, the "non-cooperative" behavior of users aggravates the congestion problem. For example, in the case of urban transport networks, travelers select their routes from an origin to a destination so as to minimize their own travel cost or travel time, which although optimal from a user's perspective (user optimization) may not be optimal from a societal one (system optimization) where a decision-maker or central controller has control of the flows on the network and seeks to allocate the flows so as to minimize the total cost in the network. Hence, before making any policy decisions on transport networks one needs to identify the underlying behavioral mechanisms regarding route selection.

This point is richly illustrated through the famous Braess (1968) paradox example, in which it is assumed that the underlying behavioral principle is that of user optimization, and travelers select their routes accordingly. In the Braess network, the addition of a new road with no change in travel demand results in all travelers in the network incurring a higher travel cost. Hence, they are all worse off after the addition of the new road. Actual practical instances of such a phenomenon have been identified in New York City and in Stuttgart, Germany. In 1990, 42nd Street in New York was closed for Earth Day, and the traffic flow in the area improved (Kolata, 1990). In Stuttgart, in turn, a new road was added to the downtown, but the traffic flow worsened and, following complaints, the new road was torn down (Bass, 1992). Interestingly, this phenomenon is also relevant to telecommunications networks (Korilis et al., 1999) and, specifically, to the Internet (Cohen and Kelly, 1990).

The coupling of transportation networks with telecommunication networks through electronic commerce, notably through business-to-business and business-to-consumer commerce, and through intelligent transportation systems is further transforming the economic landscape and affecting the movement of people, goods, and services, as well as information (Nagurney and Dong, 2002a). Telecommunication networks are assuming many of the characteristics of transport networks, including large size, non-cooperative behavior of the users, as well as congestion. In fact, telecommunication networks, in a sense, may be interpreted as transport networks in which the flows correspond to information (rather than vehicles, etc.).

In this chapter the foundations of the equilibration of transport networks will be recalled, and the evolution of modeling frameworks traced for study. The exposition is meant to be accessible to practitioners and to students, as well as to researchers in transport and to those interested in related network topics. Technical derivations and further supporting documentation are referred to in the citations. Further useful material and a supplementary chronological perspective of developments on this topic can be found in the review articles of Florian (1986) and Boyce et al. (1988), in books by Beckmann et al. (1956), Ran and Boyce (1996), Nagurney (1999, 2000), and Nagurney and Dong (2002a), and in the volumes edited by Florian (1984), Volmuller and Hamerslag (1984), Marcotte and Nguyen (1998), and Taylor (2002).

## 2. Basic decision-making concepts and models

Half a century ago, Wardrop (1952) explicitly recognized alternative possible behaviors of users of transport networks, notably, urban transport networks and stated two principles, which are commonly 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 whereas 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, as had Samuelson (1952) in the framework of spatial price equilibrium problems in which there were, however, no congestion effects. Specifically, Beckmann et al. (1956) established the equivalence between the traffic network equilibrium conditions, which state that all used paths connecting an 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 approach made the formulation, analysis, and subsequent computation of solutions to traffic network problems based on actual transportation networks realizable.

Dafermos and Sparrow (1969) coined the terms "user optimized" (U-O) and "system optimized" (S-O) for transportation networks, to distinguish between two distinct situations in which, respectively, users act unilaterally, in their own self-

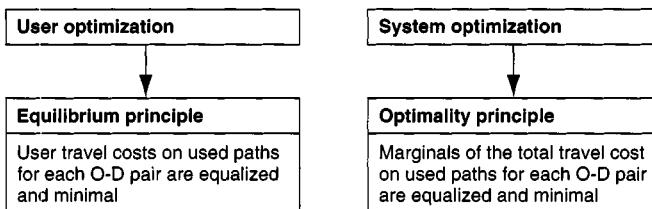


Figure 1. Distinct behavior on transportation networks.

interest, in selecting their routes, and in which users select routes according to what is optimal from a societal point of view, in that the total cost in the system is minimized. In the latter problem, marginal total costs rather than average costs are equilibrated. The former problem coincides with Wardrop's first principle, and the latter with Wardrop's second principle.

Figure 1 illustrates the two distinct behavioral principles underlying transportation networks. The concept of "system optimization" is also relevant to other types of "routing models" in transportation, as well as in communications (Bertsekas and Gallager, 1992), including those concerned with the routing of freight and computer messages, respectively. Dafermos and Sparrow (1969) also provided explicit computational procedures, i.e. 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.

## 2.1. System optimization versus user optimization

In this section, the basic transport network models are first reviewed, under distinct assumptions of their operation and distinct behavior of the users of the network. In subsequent sections, more general models are presented in which the user link cost functions are no longer separable and are also asymmetric. For such models the variational inequality formulations of the governing equilibrium conditions are also provided, since, in such cases, the conditions can no longer be reformulated as the Kuhn–Tucker conditions of a convex optimization problem.

For definiteness, and for easy reference, the classical S-O network model is presented first, followed by the classical U-O network model.

### *The S-O 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. Let  $a$  denote a link of the network connecting a pair of nodes, and let  $p$  denote a path consisting of a sequence of links connecting an O-D

pair. In transport 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 telecommunication context, however, nodes can correspond to switches or to computers, and links to telephone lines, cables, microwave links, etc. Note that paths, rather than routes, are considered here, since the former subsumes the latter. Moreover, the network concepts presented here are sufficiently general to abstract not only transport decision-making but also combined location–transport decision-making, which are returned to later. In addition, in the setting of supernetworks (Nagurney and Dong, 2002a), a path is viewed more broadly and need not be limited to a route-type decision but may, in fact, correspond to not only transport but also to telecommunications decision-making, or a combination thereof, as in the case of teleshopping and/or telecommuting.

Let  $P_\omega$  denote the set of paths connecting the O-D pair of nodes  $\omega$ . Let  $P$  denote the set of all paths in the network, and assume that there are  $J$  O-D pairs of nodes in the set  $\Omega$ . Let  $x_p$  represent the flow on path  $p$  and let  $f_a$  denote the flow on link  $a$ . The path flows on the network are grouped into the column 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 column vector  $f \in R_+^n$ , where  $n$  denotes the number of links in the network.

The following conservation of flow equation must hold:

$$f_a = \sum_{p \in P} x_p \delta_{ap}, \quad \forall a \in \mathcal{L}, \quad (1)$$

where  $\delta_{ap} = 1$  if link  $a$  is contained in path  $p$ , and 0 otherwise. Expression (1) states that the flow on a link  $a$  is equal to the sum of all the path flows on paths  $p$  that contain (traverse) link  $a$ .

Moreover, if one lets  $d_\omega$  denote the demand associated with O-D pair  $\omega$ , then one must have

$$d_\omega = \sum_{p \in P_\omega} x_p, \quad \forall \omega \in \Omega, \quad (2)$$

where  $x_p \geq 0, \forall p \in P$ ; i.e. the sum of all the path flows between an O-D pair  $\omega$  must be equal to the given demand  $d_\omega$ .

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

$$c_a = c_a(f_a), \quad \forall a \in \mathcal{L}, \quad (3)$$

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.

The total cost on link  $a$ , denoted by  $\hat{c}_a(f_a)$ , is hence given by

$$\hat{c}_a(f_a) = c_a(f_a) \times f_a, \quad \forall a \in \mathcal{L}, \quad (4)$$

i.e. the total cost on a link is equal to the user link cost on the link times the flow on the link. Here the cost is interpreted in a general sense. From a transportation engineering perspective, however, the cost on a link is assumed to typically coincide with the travel time on a link.

As noted earlier, in the S-O problem, there exists a central controller who seeks to minimize the total cost in the network system, where the total cost is expressed as

$$\sum_{a \in \mathcal{L}} \hat{c}_a(f_a), \quad (5)$$

and the total cost on a link is given by expression (4).

The S-O problem is, thus, given by

$$\text{Minimize} \quad \sum_{a \in \mathcal{L}} \hat{c}_a(f_a), \quad (6)$$

subject to

$$\sum_{p \in P_\omega} x_p = d_\omega, \quad \forall \omega \in \Omega, \quad (7)$$

$$f_a = \sum_{p \in P} x_p, \quad \forall a \in \mathcal{L}, \quad (8)$$

$$x_p \geq 0, \quad \forall p \in P. \quad (9)$$

Constraints (7) and (8), along with (9), are commonly referred to in network terminology as conservation of flow equations. In particular, they guarantee that the flow on the network, i.e. the users (travelers or computer messages, for example) do not “disappear from the network,” and, hence, are “conserved.”

The total cost on a path, denoted by  $\hat{C}_p$ , is the user cost on a path times the flow on a path, i.e.

$$\hat{C}_p = C_p x_p, \quad \forall p \in P, \quad (10)$$

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, i.e.

$$C_p = \sum_{a \in \mathcal{L}} c_a(f_a) \delta_{ap}, \quad \forall a \in \mathcal{L}. \quad (11)$$

In view of eq. (8), 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 S-O problem can be stated in path flow variables only, where one has now the problem

$$\text{Minimize} \quad \sum_{p \in P} C_p(x) x_p, \quad (12)$$

subject to constraints (7) and (9).

*System optimality conditions.* Under the assumption of increasing user link cost functions, the objective function in the S-O problem is convex, and the feasible set consisting of the linear constraints is also convex. Therefore, the optimality conditions, i.e. 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 link flow pattern  $f$ ), satisfying constraints (7)–(9) must satisfy

$$\hat{C}'_p = \begin{cases} = \mu_\omega, & \text{if } x_p > 0, \\ \geq \mu_\omega, & \text{if } x_p = 0, \end{cases} \quad (13)$$

where  $\hat{C}'_p$  denotes the marginal of the total cost on path  $p$ , given by

$$\hat{C}'_p = \sum_{a \in \mathcal{L}} \frac{\partial \hat{c}_a(f_a)}{\partial f_a} \delta_{ap}, \quad (14)$$

evaluated under conditions (13) at the solution, and  $\mu_\omega$  is the Lagrange multiplier associated with constraint (7) for that O-D pair  $\omega$ .

Observe that conditions (13) may be rewritten so that there exists an ordering of the paths for each O-D pair whereby all used paths (i.e. those with positive flow) have equal and minimal marginal total costs and the unused paths (i.e. those with zero flow) have higher (or equal) marginal total costs than those of the used paths. Hence, in the S-O problem, according to the optimality conditions (13), it is the marginal of the total cost on each used path connecting an O-D pair which is equalized and minimal (Dafermos and Sparrow, 1969).

### The U-O problem

The U-O network problem, also commonly referred to in the transportation literature as the traffic assignment problem or the traffic network equilibrium problem, is now described. Again, as in the S-O problem described above, the network  $\mathcal{G} = [\mathcal{N}, \mathcal{L}]$ , the demands associated with the O-D pairs, as well as the user link cost functions are assumed as given. Recall that U-O follows Wardrop's first principle.

*Network equilibrium conditions.* In the case of the U-O problem one seeks to determine the path flow pattern  $x^*$  (and the link flow pattern  $f^*$ ) that satisfies the conservation of flow equations (7) and (8), and the non-negativity assumption on the path flows (9), and which also satisfies the network equilibrium conditions given by the following statement. For each O-D pair  $\omega \in \Omega$  and each path  $p \in P_\omega$ :

$$C_p \begin{cases} = \lambda_\omega, & \text{if } x_p^* > 0, \\ \geq \lambda_\omega, & \text{if } x_p^* = 0. \end{cases} \quad (15)$$

Hence, in the U-O problem there is no explicit optimization concept, since users of the transport network system now act independently, in a non-cooperative manner, until they cannot improve on their situations unilaterally and, thus, an equilibrium is achieved, governed by the above equilibrium conditions. Indeed, conditions (15) are simply a restatement of Wardrop's first principle mathematically, and mean that only those paths connecting an O-D pair will be used that have equal and minimal user costs. In conditions (15) the minimal cost for a given O-D pair is denoted by  $\lambda_\omega$  and its value is obtained once the equilibrium flow pattern is determined.

Otherwise, a user of the network could improve upon his or her situation by switching to a path with lower cost. User optimization represents decentralized decision-making, whereas S-O represents centralized decision-making. (See also Figure 1.)

In order to obtain a solution to the above problem, Beckmann et al. (1956) established that the solution to the equilibrium problem, in the case of user link cost functions (cf. eq. (3)) in which the cost on a link only depends on the flow on that link could be obtained by solving the following optimization problem:

$$\text{Minimize} \quad \sum_{a \in \mathcal{L}} \int_0^{f_a} c_a(y) dy, \quad (16)$$

subject to

$$\sum_{p \in P_\omega} x_p = d_\omega, \quad \forall \omega \in \Omega, \quad (17)$$

$$f_a = \sum_{p \in P} x_p \delta_{ap}, \quad \forall a \in \mathcal{L}, \quad (18)$$

$$x_p \geq 0, \quad \forall p \in P. \quad (19)$$

Note that the conservation of flow equations are identical in both the U-O network problem (see conditions (17)–(19)) and the S-O problem (see conditions (7)–(9)). The behavior of the individual decision-makers, termed “users” is, however, different. Users of the network system, who generate the flow on the network, now act independently, and are not controlled by a centralized controller.

The objective function given by expression (16) 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 S-O problem given by expression (6), or, equivalently, by expression (12). Note that in the case of separable, as well as non-separable but symmetric (which will be returned to later), user link cost functions the  $\lambda_\omega$  term in conditions (15) corresponds to the Lagrange multiplier associated with the constraint (17) for that O-D pair  $\omega$ . However, in the case of non-separable and asymmetric functions there is no optimization reformulation of the traffic network equilibrium conditions (15), and the  $\lambda_\omega$  term simply reflects the minimum user cost associated with the O-D pair  $\omega$  at equilibrium.

### 3. Models with asymmetric link costs

There has been much research activity in the past several decades in terms of both the modeling and the development of methodologies to enable the formulation and computation of more general traffic (and related) network equilibrium models. Examples of general models include those that allow for multiple modes of transportation or multiple classes of users, who perceive cost on a link in an individual way. In this section, network models in which the user cost on a link is no longer dependent solely on the flow on that link are considered.

Assume that user link cost functions are now of a general form, i.e. the cost on a link may depend not only on the flow on the link but on other link flows on the network:

$$c_a = c_a(f), \quad \forall a \in \mathcal{L}. \quad (20)$$

In the case where the symmetry assumption exists, i.e.  $\partial c_a(f)/\partial f_b = \partial c_b(f)/\partial f_a$ , for all links  $a, b \in \mathcal{L}$ , one can still reformulate the solution to the network equilibrium problem satisfying equilibrium conditions (15) as the solution to an optimization problem (Dafermos, 1972), 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. Models of traffic networks with asymmetric cost functions are important since they allow for the formulation, qualitative analysis, and, ultimately, given the state of the art, 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 the flow on that link. Such a generalization allows for a more realistic treatment of intersections, two-way links, multiple modes of transport as well as distinct classes of users of the network.

Indeed, it was in the domain of such traffic network equilibrium problems that the theory of finite-dimensional variational inequalities realized its earliest

success, beginning with the contributions of Dafermos (1980). Nagurney (1999) provides an introduction to the subject, as well as describing applications ranging from traffic network and spatial price equilibrium problems to financial equilibrium problems. Variational inequality formulations of both fixed-demand and elastic-demand traffic network equilibrium problems are presented below.

The S-O problem, in turn, in the case of non-separable user link cost functions becomes (see also expressions (6)–(9))

$$\text{Minimize} \quad \sum_{a \in \mathcal{L}} \hat{c}_a(f), \quad (21)$$

subject to conditions (7)–(9), where  $\hat{c}_a(f) = c_a(f) \times f_a, \forall a \in \mathcal{L}$ .

The system optimality conditions remain as in expressions (13), but now the marginal of the total cost on a path becomes, in this more general case,

$$\hat{C}'_p = \sum_{a, b \in \mathcal{L}} \frac{\partial \hat{c}_b(f)}{\partial f_a} \delta_{ap}, \quad \forall p \in P. \quad (22)$$

*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, i.e. to the network equilibrium, problem using standard optimization algorithms. It is emphasized, 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 Nagurney and Dong (2002a). The variational inequality of the network equilibrium conditions in path flows as well as in link flows will be provided.

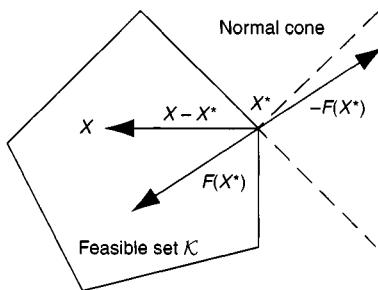
Specifically, the variational inequality problem (finite-dimensional) is defined as follows:

*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

$$\langle F(X^*), X - X^* \rangle \geq 0, \quad \forall X \in \mathcal{K}, \quad (23)$$

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  $R^N$ .

Figure 2. Geometric interpretation of  $\text{VI}(F, \mathcal{K})$ .

Variational inequality (23) 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.

A geometric interpretation of the variational inequality problem  $\text{VI}(F, \mathcal{K})$  is given in Figure 2. In particular,  $F(X^*)$  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, i.e. it satisfies equilibrium conditions (15) if and only if it satisfies the variational inequality problem

$$\sum_{\omega \in \Omega} \sum_{p \in P_\omega} C_p(x^*) \times (x - x^*) \geq 0, \quad \forall x \in K^1, \quad (24)$$

or, in vector form,

$$\langle C(x^*), x - x^* \rangle \geq 0, \quad \forall x \in K^1, \quad (25)$$

where  $C$  is the  $n_p$ -dimensional column vector of path user costs, and  $K^1$  is defined as  $K^1 \equiv \{x \geq 0, \text{ such that eq. (17) holds}\}$ .

*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

$$\sum_{a \in \mathcal{L}} c_a(f^*) \times (f_a - f_a^*) \geq 0, \quad \forall f \in K^2, \quad (26)$$

or, in vector form,

$$\langle c(f^*), f - f^* \rangle \geq 0, \quad \forall f \in K^2, \quad (27)$$

where  $c$  is the  $n$ -dimensional column vector of link user costs, and  $K^2$  is defined as  $K^2 \equiv \{f \mid \text{there exists an } x \geq 0 \text{ and satisfying eqs (17) and (18)}\}$ .

Note that one may put variational inequality (25) into standard form (23) by letting  $F \equiv C$ ,  $X \equiv x$ , and  $\mathcal{K} \equiv K^1$ . Also, one may put variational inequality (27) into standard form where now  $F \equiv c$ ,  $X \equiv f$ , and  $\mathcal{K} \equiv K^2$ .

Alternative variational inequality formulations of a problem are useful for devising other models, including dynamic versions, as well as for purposes of computation using different algorithms. In Section 5, the relationship between variational inequality formulations and projected dynamical systems, in which the latter provides the disequilibrium dynamics prior to the attainment of the equilibrium, as formulated via the former, is described.

The theory of variational inequalities (Kinderlehrer and Stampacchia, 1980) allows one to qualitatively analyze the equilibrium patterns in terms of existence and uniqueness, as well as of the 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, including the aforementioned equilibration algorithms of Dafermos and Sparrow (1969), which exploit network structure as well as the commonly used in practice Frank-Wolfe (1956) algorithm. In particular, projection methods as well as relaxation methods (Dafermos, 1980, 1982; Florian and Spiess, 1982; Nagurney, 1999) have been successfully applied to compute solutions to variational inequality formulations of traffic network equilibrium problems.

It should be emphasized 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. This was recognized by Dafermos in 1976 (in the context of separable link cost functions) in her development of such integrated traffic network equilibrium models in which location decisions are made simultaneous to transportation route decisions.

It is worth noting that the presentation of the variational inequality formulations of the fixed-demand models given above was in the context of single-mode (or single-class) transport networks. It should be emphasized, however, that

in view of the generality of the functions considered (cf. expression (20)), the modeling framework described above can also be adapted to multimodal/multiclass problems in which there are multiple modes of transport available and/or multiple classes of users, each of whom perceives the cost on the links of the network in an individual manner. Dafermos, in 1972, demonstrated how, through a formal model, a multiclass traffic network could be cast into a single-class network through the construction of an expanded (and, again, abstract) network consisting of as many copies of the original network as there were classes. The application of such a transformation is also relevant to telecommunication networks.

Also, note that here the focus is on deterministic network equilibrium problems. Some basic stochastic traffic network equilibrium models can be found in Sheffi (1985). Dial (1971) is credited with developing the first stochastic route choice model. Daganzo and Sheffi (1977), in turn, formulated a stochastic U-O traffic network model with route choice in which the equilibrium criterion could be succinctly stated, as no traveler can improve his or her perceived travel time by unilaterally changing routes.

Finally, it should be emphasized that the dynamic models (although presented in a deterministic framework) have been analyzed qualitatively using tools from stochastic processes (Dupuis and Nagurney, 1993; Nagurney and Zhang, 1996).

*Variational inequality formulations of elastic-demand problems.* A general network equilibrium model with elastic demands due to Dafermos (1982) will now be described. Specifically, it is assumed that one has associated with each O-D pair  $\omega$  in the network a travel disutility function  $\lambda_\omega$ , 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, i.e.

$$\lambda_\omega = \lambda_\omega(d), \quad \forall \omega \in \Omega, \quad (28)$$

where  $d$  is the  $J$ -dimensional column vector of the demands.

The notation, otherwise, is as described earlier, except that, here, user link cost functions that are general, i.e. of the form (20), are also considered. The conservation of flow equations (see also (1) and (2)), in turn, are given by

$$f_a = \sum_{p \in P} x_p \delta_{ap}, \quad \forall a \in \mathcal{L}, \quad (29)$$

$$d_\omega = \sum_{p \in P_\omega} x_p, \quad \forall \omega \in \Omega, \quad (30)$$

$$x_p \geq 0, \quad \forall p \in P. \quad (31)$$

Hence, in the elastic-demand case, the demands in expression (30) are now variables and no longer given, as was the case for the fixed-demand expression (2).

*Network equilibrium conditions in the case of elastic demand.* The network equilibrium conditions (see also conditions (15)) now 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 (30) and (31) (which induces a link flow pattern  $f^*$  through eq. (29)) is a network equilibrium pattern if it satisfies

$$C_p(x^*) \begin{cases} = \lambda_\omega(d^*), & \text{if } x_p^* > 0, \\ \geq \lambda_\omega(d^*), & \text{if } x_p^* = 0. \end{cases} \quad (32)$$

Equilibrium conditions (32) 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. Conditions (32) can be given an economic interpretation, as described in Section 1. 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. It should be emphasized that 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 also 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.

Also, note that although the presentation of the elastic-demand traffic network model has been in the case of a single mode of transport or class of user one can readily (with an accompanying increase in notation) explicitly introduce distinct modes to the above model as follows. One needs only to introduce subscripts to denote modes/classes, redefine all of the above vectors accordingly, and the conservation of flow equations, and state that conditions (32) then must hold for each mode/class. In other words, in equilibrium, the used paths for a given mode and O-D pair must have minimal and equal user path costs, which in turn, must be equal to the travel disutility for that mode and O-D pair at the equilibrium demand. Of course, as described in the case of fixed demands, one can also have made as many copies as there are modes on the network, in which case the above single-modal but extended elastic-demand model would be equivalent to the multimodal one.

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 (32) are presented. These are analogues of the formulations (24) and (25), and (26) and (27), respectively, for the fixed-demand model.

*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, i.e. it satisfies equilibrium conditions (32) if and only if it satisfies the variational inequality problem

$$\sum_{\omega \in \Omega} \sum_{p \in P_\omega} C_p(x^*) \times (x - x^*) - \sum_{\omega \in \Omega} \lambda_\omega(d^*) \times (d_\omega - d_\omega^*) \geq 0, \quad \forall (x, d) \in K^3, \quad (33)$$

or, in vector form,

$$\langle C(x^*), x - x^* \rangle - \langle \lambda(d^*), d - d^* \rangle \geq 0, \quad \forall (x, d) \in K^3, \quad (34)$$

where  $\lambda$  is the  $J$ -dimensional vector of disutilities, and  $K^3$  is defined as:  $K^3 \equiv \{x \geq 0, \text{ such that eq. (30) holds}\}$ .

*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

$$\sum_{a \in \mathcal{L}} c_a(f^*) \times (f_a - f_a^*) - \sum_{\omega \in \Omega} \lambda_\omega(d^*) \times (d_\omega - d_\omega^*) \geq 0, \quad \forall (f, d) \in K^4, \quad (35)$$

or, in vector form,

$$\langle c(f^*), f - f^* \rangle - \langle \lambda(d^*), d - d^* \rangle \geq 0, \quad \forall (f, d) \in K^4, \quad (36)$$

where  $K^4 \equiv \{(f, d), \text{ such that there exists an } x \geq 0 \text{ satisfying eqs (17) and (18)}\}$ .

Note that, under the symmetry assumption on the disutility functions, i.e. if  $\partial \lambda_\omega / \partial d_\omega = \partial \lambda_\omega / \partial d_w$ , for all  $w, \omega$ , in addition to such an assumption on the user link cost functions (see the text following eq. (20)), one can obtain an optimization reformulation of the network equilibrium conditions (32), which in the case of separable user link cost functions and disutility functions is given by

$$\text{Minimize} \quad \sum_{a \in \mathcal{L}} \int_0^{f_a} c_a(y) dy - \sum_{\omega \in \Omega} \int_0^{d_\omega} \lambda_\omega(z) dz, \quad (37)$$

subject to conditions (29)–(31).

It should be noted that the elastic-demand model described above is related closely to the well-known spatial price equilibrium models of Samuelson (1952) and Florian and Los (1982). Indeed, as demonstrated by Dafermos and Nagurney (1985) in the context of a single commodity, and, subsequently, by Dafermos (1986) in the case of multiple commodities, spatial price equilibrium problems are isomorphic to traffic network equilibrium problems over appropriately constructed networks. Hence, the well-developed theory of traffic networks can be transferred to the study of commodity flows in the case of spatial price equilibrium in which the equilibrium production, consumption, and commodity trade flows are to be determined satisfying the equilibrium conditions that there will be a positive flow (in equilibrium) of the commodity between a pair of supply and demand markets if the supply price at the supply market plus the unit cost of transportation is equal to the demand price at the demand market. A variety of such models and associated references can be found in Nagurney (1999) and Nagurney and Zhang (1996).

Although the focus of this chapter is on transport network equilibrium models in an urban setting, models of freight networks are closely related to those discussed above. Of course, one must distinguish the behavior of the operators of such networks and model the competition accordingly (Friesz and Harker, 1985).

An example of a simple elastic-demand network equilibrium problem is now presented, to conclude this section.

### *Example*

Consider the network depicted in Figure 3 in which there are three nodes: 1, 2, and 3; three links:  $a$ ,  $b$ , and  $c$ ; and a single O-D pair  $\omega_1 = (1, 3)$ . Let path  $p_1 = (a, c)$  and path  $p_2 = (b, c)$ .

Assume that the user link cost functions are

$$c_a(f) = 5f_a + 2f_b + 15, \quad c_b(f) = 7f_b + f_a + 15, \quad c_c(f) = 3f_c + f_a + f_b + 12,$$

and the disutility (or inverse demand) function is given by

$$\lambda_{\omega_1}(d_{\omega_1}) = -2d_{\omega_1} + 114.$$

Observe that in this example, the user link cost functions are non-separable and asymmetric and, hence, the equilibrium conditions (32) cannot be reformulated as the solution to an optimization problem, but, rather, as the solution to the variational inequalities (33) (or (34)) or (35) (or (36)).

The U-O flow and demand pattern that satisfies equilibrium conditions (32) is  $x_{p_1}^* = 5$ ,  $x_{p_2}^* = 4$ , and  $d_{\omega_1}^* = 9$ , with the associated link flow pattern  $f_a^* = 5$ ,  $f_b^* = 4$ , and  $f_c^* = 9$ .

The incurred user costs on the paths are  $C_{p_1} = C_{p_2} = 96$ , which is precisely the value of the disutility  $\lambda_{\omega_1}$ . Hence, this flow and demand pattern satisfies

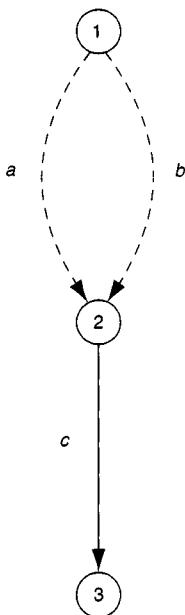


Figure 3. An elastic-demand example.

equilibrium conditions (32). Indeed, both paths  $p_1$  and  $p_2$  are utilized, and their user paths costs are equal to each other. In addition, these costs are equal to the disutility associated with the O-D pair that the two paths connect.

#### 4. Multiclass, multicriteria traffic network equilibrium models

In this section, multiclass, multicriteria network equilibrium models that can serve as alternatives to multimodal traffic network equilibrium models are described. These models are important since they allow for the individual weighting of distinct criteria associated with decision-making on networks and, especially, transport networks. Moreover, such models have been successful in formalizing decision-making surrounding transport/telecommunication network trade-offs, as in the case of telecommuting versus commuting decision-making, and tele-shopping versus shopping decision-making (Nagurney and Dong, 2002a,b,c; Nagurney et al., 2002a,b).

Multicriteria traffic network models were introduced by Quandt (1967) and Schneider (1968), and explicitly consider that travelers may be faced with several

criteria, notably travel time and travel cost, when selecting their optimal routes of travel. The ideas were further developed by Dial (1979), who proposed an uncongested model, and Dafermos (1981), who introduced congestion effects and derived an infinite-dimensional variational inequality formulation of her multiclass, multicriteria traffic network equilibrium problem, along with some qualitative properties. The paper by Nagurney and Dong (2002b) provides a chronology of citations that highlights the number of criteria treated by various authors, typically travel time and travel cost; whether or not these functions are allowed to be flow-dependent or not; and the form (separable or general) handled. In addition, it notes the type of demand considered, i.e. fixed or elastic, and whether the demand is class-dependent, and, if elastic, what form the demand functions take. Moreover, it provides the type of methodology used in the formulation and analysis such as, for example, an optimization approach, a finite-dimensional variational inequality approach, or an infinite-dimensional approach, along with whether the citation contains algorithmic contributions and/or qualitative ones. Note that, in the case of infinite-dimensional variational inequality formulations, the number of classes is, usually, infinite, whereas in the case of finite-dimensional formulations, the number of classes is assumed to be finite. Additional citations, including literature exploring multicriteria traffic models used in practice, may be found in Leurent (1998).

In this section, for completeness, recall the multiclass, multicriteria network equilibrium model with elastic demand developed by Nagurney and Dong (2002b). The model has the following novel and significant features:

- It includes weights associated with the two criteria of travel time and travel cost, which are not only class-dependent but also, explicitly, link-dependent. Such weights may incorporate such subjective factors as the relative safety or risk associated with particular links, the relative comfort, or even the view.
- It treats demand functions (rather than their inverses), which are very general and not separable functions. Specifically, the demand associated with a class and O-D pair can depend not only on the travel disutility of different classes traveling between the particular O-D pair but can also be influenced by the disutilities of the classes traveling between other O-D pairs. Hence, the model has implications for locational choice (Boyce et al., 1983).

As in the transport network models described in Sections 2 and 3, a general network  $\mathcal{G} = [\mathcal{N}, \mathcal{L}]$ , where  $\mathcal{N}$  denotes the set of nodes in the network and  $\mathcal{L}$  the set of directed links, is considered. Let  $a$  denote a link of the network connecting a pair of nodes and let  $p$  denote a path, assumed to be acyclic, consisting of a sequence of links connecting an O-D pair of nodes. There are  $n$  links in the network and  $n_p$  paths. Let  $\Omega$  denote the set of  $J$  O-D pairs. The set of paths

connecting the O-D pair  $\omega$  is denoted by  $P_\omega$ , and the entire set of paths in the network by  $P$ .

Assume now that there are  $k$  classes of travelers in the network with a typical class denoted by  $i$ . Let  $f_a^i$  denote the flow of class  $i$  on link  $a$ , and let  $x_p^i$  denote the non-negative flow of class  $i$  on path  $p$ . The relationship between the link flows by class and the path flows is then

$$f_a^i = \sum_{p \in P} x_p^i \delta_{ap}, \quad \forall i, \forall a, \quad (38)$$

where  $\delta_{ap} = 1$  if link  $a$  is contained in path  $p$ , and 0 otherwise. Hence, the flow of a class of traveler on a link is equal to the sum of the flows of the class on the paths that contain that link.

In addition, let  $f_a$  now denote the total flow on link  $a$ , where

$$f_a = \sum_{i=1}^k f_a^i, \quad \forall a \in \mathcal{L}. \quad (39)$$

Group the class link flows into the  $kn$ -dimensional column vector  $\tilde{f}$  with components  $\{f_a^1, \dots, f_a^1, \dots, f_a^k, \dots, f_a^k\}$  and the total link flows  $\{f_a, \dots, f_a\}$  into the  $n$ -dimensional column vector  $f$ . Also, group the class path flows into the  $kn_p$ -dimensional column vector  $\tilde{x}$  with components  $\{x_{p_1}^1, \dots, x_{p_{n_p}}^k\}$ .

The functions associated with the links can now be described. Assume, as given, a travel time function  $t_a$  associated with each link  $a$  in the network, where

$$t_a = t_a(f), \quad \forall a \in \mathcal{L}, \quad (40)$$

and a travel cost function  $c_a$  associated with each link  $a$ , i.e.

$$c_a = c_a(f), \quad \forall a \in \mathcal{L}, \quad (41)$$

with both these functions assumed to be continuous. Note that allowance is made here for the general situation in which both the travel time and the travel cost can depend on the entire link flow pattern, whereas in Dafermos (1981) it was assumed that these functions were separable.

Assume that each class of traveler  $i$  has his or her own perception of the trade-off between travel time and travel cost, represented by the non-negative weights  $w_{1a}^i$  and  $w_{2a}^i$ . Here,  $w_{1a}^i$  denotes the weight associated with the travel time of class  $i$  on link  $a$ , and  $w_{2a}^i$  denotes the weight associated with the travel cost of class  $i$  on link  $a$ . The weights  $w_{1a}^i$  and  $w_{2a}^i$  are link-dependent and, hence, can incorporate such link-dependent factors as safety, comfort, and view. For example, in the case of a pleasant view on a link, travelers may weight the travel cost higher than the travel time on such a link. However, if a link has a rough surface or is noted for unsafe road conditions such as ice in the winter, travelers may then assign a higher weight to the travel time than to the travel cost. Link-dependent weights provide a

greater level of generality and flexibility in modeling travel decision-making than weights that are identical for the travel time and for the travel cost on all links for a given class.

Then, construct the generalized cost/disutility of class  $i$  associated with link  $a$ , and denoted by  $u_a^i$ , as

$$u_a^i = w_{1a}^i t_a + w_{2a}^i c_a, \quad \forall i, \forall a. \quad (42)$$

In view of expressions (39), (40), and (41),

$$u_a^i = u_a^i(\tilde{f}), \quad \forall i, \forall a, \quad (43)$$

can be written, and the link generalized costs grouped into the  $kn$ -dimensional column vector  $u$  with components  $\{u_a^1, \dots, u_n^1, \dots, u_a^k, \dots, u_n^k\}$ .

Observe that a possible weighting scheme would be  $w_{1a}^i = \psi_a^i$  and  $w_{2a}^i = (1 - \psi_a^i)$ , with  $\psi_a^i$  lying in the range from 0 to 1 with  $\psi_a^i = 1$  denoting a class of traveler concerned only with the travel time on a particular link  $a$ , and with  $\psi_a^i = 0$  denoting a class of traveler concerned only about travel cost on link  $a$ ; with weights within the range reflecting classes who perceive travel time and travel cost as per the disutility functions accordingly. Dafermos (1981) proposed such a weighting scheme in which  $w_{1a}^i = \psi^i$  and  $w_{2a}^i = (1 - \psi^i)$  for all links  $a$  and classes  $i$ . Such a weighting scheme has an interpretation of a weighted average, but is not link-dependent.

Let  $v_p^i$  denote the generalized cost of class  $i$  associated with traveling on path  $p$ , where

$$v_p^i = \sum_{a \in \mathcal{L}} u_a^i(\tilde{f}) \delta_{ap}, \quad \forall i, \forall p. \quad (44)$$

Hence, the generalized cost, as perceived by a class, associated with traveling on a path is the sum of the generalized link costs on links comprising the path.

Let  $d_\omega^i$  denote the travel demand of class  $i$  traveler between O-D pair  $\omega$ , and let  $\lambda_\omega^i$  denote the travel disutility associated with a class  $i$  traveler traveling between the O-D pair  $\omega$ . The travel demands are then grouped into a  $kJ$ -dimensional column vector  $d$  and the O-D pair travel disutilities into a  $kJ$ -dimensional column vector  $\lambda$ .

The path flow vector  $\tilde{x}$  induces the demand vector  $d$  with components

$$d_\omega^i = \sum_{p \in P_\omega} x_p^i, \quad \forall i, \forall \omega. \quad (45)$$

Assume that the travel demands are determined by the O-D travel disutilities, i.e.

$$d_\omega^i = d_\omega^i(\lambda), \quad \forall i, \forall \omega, \quad (46)$$

and denote the  $kJ$ -dimensional row vector of demand functions by  $d(\lambda)$ .

Note that the travel demand function (46) is quite general and has choice location implications as well. For example, it allows the demand for a class associated with an O-D pair to depend not only on the travel disutilities of different classes associated with that O-D pair but also on those associated with other O-D pairs.

#### 4.1. Traffic network equilibrium conditions

The traffic network equilibrium conditions in the case of elastic-travel demands (Dafermos and Nagurney, 1984), in the generalized context of the multiclass, multicriteria traffic network equilibrium problem, take on the following form. For each class  $i$ , for all O-D pairs  $\omega \in \Omega$ , and for all paths  $p \in P_\omega$ , the flow pattern  $\tilde{x}$  is said to be in equilibrium if the following conditions hold:

$$v_p^i(\tilde{f}^*) \begin{cases} = \lambda_\omega^{i*}, & \text{if } x_p^{i*} > 0, \\ \geq \lambda_\omega^{i*}, & \text{if } x_p^{i*} = 0, \end{cases} \quad (47)$$

and

$$d_\omega^{i*}(\lambda^*) \begin{cases} = \sum_{p \in P_\omega} x_p^{i*}, & \text{if } \lambda_\omega^{i*} > 0, \\ \leq \sum_{p \in P_\omega} x_p^{i*}, & \text{if } \lambda_\omega^{i*} = 0. \end{cases} \quad (48)$$

In other words, all utilized paths by a class connecting an O-D pair have equal and minimal generalized path costs. Meanwhile, if the travel disutility associated with traveling between O-D pair  $\omega$  of class  $i$  is positive, then the market clears for that O-D pair and that class, i.e. the sum of the path flows of that class of traveler on paths connecting that O-D pair is equal to the demand associated with that O-D pair; if the travel disutility is zero, then the sum of the path flows can exceed the demand of that class of traveler.

Hence, in the elastic-demand framework, different classes of travelers can also choose their O-D pairs, in addition to their paths. Thus, this model allows one to capture the relative attractiveness of different O-D pairs as perceived by the distinct classes of travelers through the travel disutilities.

The feasible set  $K$  underlying the problem is defined as  $K \equiv \{(\tilde{f}, d, \lambda)/\lambda \geq 0 \text{ and } \exists \tilde{x} \geq 0, \text{ such that eqs (38), (39), and (45) hold}\}$ .

*Theorem 5: variational inequality formulation of multiclass, multicriteria network equilibrium with elastic demands – link flow version*

A multiclass, multicriteria link flow, travel demand, and O-D travel disutility pattern  $(\tilde{f}^*, d^*, \lambda^*) \in K$  is a traffic network equilibrium, i.e. it satisfies

equilibrium conditions (47) and (48) if and only if it satisfies the variational inequality problem

$$\begin{aligned} \sum_{i=1}^k \sum_{a \in \mathcal{L}} u_a^i(\tilde{f}^*) \times (f_a^i - f_a^{i*}) - \sum_{i=1}^k \sum_{\omega \in W} \lambda_\omega^{i*} \times (d_\omega^i - d_\omega^{i*}) + \\ \sum_{i=1}^k \sum_{\omega \in W} (d_\omega^{i*} - d_\omega^i(\lambda^*)) \times (\lambda_\omega^i - \lambda_\omega^{i*}) \geq 0, \quad \forall (\tilde{f}, d, \lambda) \in \mathcal{K}, \end{aligned} \quad (49)$$

or, equivalently, in standard form,

$$\langle F(X^*, X - X^*) \rangle \geq 0, \quad \forall X \in \mathcal{K}, \quad (50)$$

where  $F \equiv (u, -\lambda, d - d(\lambda))$  and  $X \equiv (\tilde{f}, d, \lambda)$ .

Nagurney and Dong (2002a) discuss network equilibrium models with multiple criteria and multiple classes in which there are a finite number of criteria associated with decision-making and the weights (as those above) are also link- and class-dependent. It is emphasized that the qualitative analysis of such models is more challenging than of those presented in the preceding sections since the criterion functions are in terms of the total links flows whereas the generalized costs are constructed according to the classes and links.

## 5. Dynamics

In this section, a brief summary is provided of how projected dynamical systems theory can be applied to the elastic-demand traffic network equilibrium problem presented in Section 3 in order to provide the disequilibrium dynamics. Dupuis and Nagurney (1993) proved that, given a variational inequality problem, there is a naturally associated dynamical system, the set of stationary points of which coincides precisely with the set of solutions of the variational inequality problem. The dynamical system, termed a projected dynamical system by Zhang and Nagurney (1995), is non-classical in that its right-hand side, which is a projection operator, is discontinuous. Nevertheless, it can be qualitatively analyzed and approximated through discrete-time algorithms as described in Nagurney and Zhang (1996). Importantly, projected dynamical systems theory provides insights into travelers' dynamic behavior in making their trip decisions and in adjusting their route choices. Moreover, it provides for a powerful theory of stability analysis. Other approaches to dynamic traffic network problems can be found in Ran and Boyce (1996). In particular, the focus here is on the disequilibrium dynamics and on what can be viewed as the day-to-day adjustment until an equilibrium is reached.

Since users on a network select paths so as to reach their destinations from their origins, consider variational inequality (34) as the basic one for the dynamical system equivalence. Specifically, note that, in view of constraint (30), one may define  $\hat{\lambda}(x) \equiv \lambda(d)$ , in which case variational inequality (30) can be rewritten in the path flow variables  $x$  only, i.e. it is sought to determine  $x^* \in R_+^{n_p}$ , such that

$$\langle C(x^*) - \bar{\lambda}(x^*), x - x^* \rangle \geq 0, \quad \forall x \in R_+^{n_p}, \quad (51)$$

where  $\bar{\lambda}(x)$  is the  $(n_{P_{\omega_1}} \times n_{P_{\omega_2}} \times \dots \times n_{P_{\omega_J}})$ -dimensional column vector with components

$$(\hat{\lambda}_{\omega_1}(x), \dots, \hat{\lambda}_{\omega_1}(x), \dots, \hat{\lambda}_{\omega_J}(x), \dots, \hat{\lambda}_{\omega_J}(x)).$$

If  $X \equiv x$  and  $F(X) \equiv C(x) - \bar{\lambda}(x)$  and  $K \equiv \{x \mid x \in R_+^{n_p}\}$ , then, clearly, expression (51) can be put into the standard form given by expression (23). The dynamical system, first presented by Dupuis and Nagurney (1993), whose stationary points correspond to solutions of expression (51), is given by

$$\dot{x} = \Pi_K(x, \bar{\lambda}(x) - C(x)), \quad x(0) = x_0 \in K, \quad (52)$$

where the projection operator  $\Pi_K(x, v)$  is defined as

$$\Pi_K(x, v) = \lim_{\delta \rightarrow 0} \frac{P_K(x + \delta v) - x}{\delta}, \quad (53)$$

and

$$P_K = \operatorname{argmin}_{z \in K} \|z - x\|. \quad (54)$$

The dynamics described by expression (52) are as follows. The rate of change of flow on a path connecting an O-D pair is equal to the difference between the travel disutility for that O-D pair and the cost on that path at that instance in time. If the path cost exceeds the travel disutility, then the flow on the path will decrease; if it is less than the disutility, then the flow on that path will increase. The projection operator in expression (30) guarantees that the flow on the paths will not be negative, since this would violate feasibility. Hence, the path flows (and incurred travel demands) evolve from an initial path flow pattern at time zero given by  $x(0)$  until a stationary point is reached, i.e. when  $\dot{x} = 0$ ; at which point, for that particular  $x^*$ ,

$$\dot{x} = 0 = \Pi_K(x^*, \bar{\lambda}(x^*) - C(x^*)), \quad (55)$$

and  $x^*$  also solves variational inequality (51), and is, hence, a traffic network equilibrium satisfying the elastic-demand equilibrium conditions (32).

Qualitative properties of the dynamic trajectories, as well as conditions for stability of the solutions as well as discrete-time algorithms can be found in Zhang and Nagurney (1995). In particular, note that discrete-time algorithms such as

those proposed in Nagurney and Zhang (1996) provide for a time discretization of the continuous time trajectories and may also be interpreted as discrete-time adjustment processes.

## 6. Summary and new directions

In this chapter the evolution of the foundations of transport network equilibrium modeling and analysis have been traced, with a focus on the principal methodological advances. In particular, an attempt has been made to set out in accessible fashion rigorous approaches to the formulation of a variety of traffic network equilibrium models and to establish relationships between the models as well as those that are closely linked such as spatial price equilibrium models.

It should be emphasized that this topic is a very active area of research as well as practice. It should also be highlighted and further emphasized that traffic network equilibrium modeling and analysis provide a powerful framework for decision-making on complex networks, in general. Indeed, given the interrelationships between telecommunication and transportation networks in today's network economy, we can expect further synergies and advances in the study of such foundational networks. Of particular promise are the areas of multicriteria decision-making on networks, multitiered networks, as well as multilevel networks (in the form of transportation/logistical/financial/informational networks), formally referred to as supernetworks. Nagurney and Dong (2002a), and the references therein, provide further background on spatial equilibration in transportation.

## Acknowledgments

The preparation of this chapter was supported, in part, by NSF Grant No. IIS-0002647 and by a 2001 AT&T Industrial Ecology Fellowship. This support is gratefully acknowledged.

## References

- Banister, D. and Button, K.J. (1993) "Environmental policy and transport: an overview," in: D. Banister and K.J. Button, eds, *Transport, the environment, and sustainable development*. London: Spon.
- Bar-Gera, H. (1999) *Origin-based algorithms for transportation network modeling*, Technical Report 103. Research Triangle Park: National Institute of Statistical Sciences.
- Bass, T. (1992) "Road to ruin," *Discover*, May:56–61.
- Beckmann, M.J., C.B. McGuire and C.B. Winsten (1956) *Studies in the economics of transportation*. New Haven: Yale University Press.
- Bertsekas, D.P. and R. Gallager (1992) *Data networks*, 2nd edn. Englewood Cliffs: Prentice-Hall.

- Boyce, D.E., K.S. Chon, Y.J. Lee, K.T. Lin and L.J. LeBlanc (1983) "Implementation and computational issues for combined models of location, destination, mode, and route choice," *Environment and Planning A*, 15:1219–1230.
- Boyce, D.E., L.J. LeBlanc and K.S. Chon (1988) "Network equilibrium models of urban location and travel choices: a retrospective survey," *Journal of Regional Science*, 28:159–183.
- Braess, D. (1968) "Über ein Paradoxon der Verkehrsplanung," *Unternehmensforschung*, 12:258–268.
- Cohen, J. and F.P. Kelly (1990) "A paradox of congestion on a queuing network," *Journal of Applied Probability*, 27:730–734.
- Dafermos, S.C. (1972) "The traffic assignment problem for multimodal networks," *Transportation Science*, 6:73–87.
- Dafermos, S.C. (1976) "Integrated equilibrium flow models for transportation planning," in: M.A. Florian, ed., *Traffic equilibrium methods. Lecture notes in economics and mathematical systems 118*. New York: Springer-Verlag.
- Dafermos, S. (1980) "Traffic equilibrium and variational inequalities," *Transportation Science*, 14:42–54.
- Dafermos, S. (1981) *A multicriteria route-mode choice traffic equilibrium model*. Providence: Lefschetz Center for Dynamical Systems, Brown University.
- Dafermos, S. (1982) "The general multimodal network equilibrium problem with elastic demand," *Networks*, 12:57–72.
- Dafermos, S. (1986) "Isomorphic multiclass spatial price and multimodal traffic network equilibrium models," *Regional Science and Urban Economics*, 16:197–209.
- Dafermos, S. and A. Nagurney (1984) "Stability and sensitivity analysis for the general network equilibrium – travel choice model," in: J. Volmuller and R. Hamerslag, eds, *Proceedings of the Ninth International Symposium on Transportation and Traffic Theory*. Utrecht: VNU Press.
- Dafermos, S. and A. Nagurney (1985) *Isomorphism between spatial price and traffic network equilibrium models*, LCDS 85–17. Providence: Lefschetz Center for Dynamical Systems, Brown University.
- Dafermos, S.C. and F.T. Sparrow (1969) "The traffic assignment problem for a general network," *Journal of Research of the National Bureau of Standards B*, 73:91–118.
- Daganzo, C.F. and Y. Sheffi (1977) "On stochastic models of traffic assignment," *Transportation Science*, 11:253–174.
- Dial, R.B. (1971) "Probabilistic multipath traffic assignment model which obviates path enumeration," *Transportation Research*, 5:83–111.
- Dial, R.B. (1979) "A model and algorithms for multicriteria route-mode choice," *Transportation Research B*, 13:311–316.
- Dupuis, P. and A. Nagurney (1993) "Dynamical systems and variational inequalities," *Annals of Operations Research*, 44:9–42.
- Florian, M., ed. (1984) *Transportation planning models*. Amsterdam: North Holland.
- Florian, M. (1986) "Nonlinear cost network models in transportation analysis," *Mathematical Programming Study*, 26:167–196.
- Florian, M. and M. Los (1982) "A new look at static spatial price equilibrium models," *Regional Science and Urban Economics*, 12:579–597.
- Florian, M. and H. Spiess (1982) "The convergence of diagonalization algorithms for asymmetric network equilibrium problems," *Transportation Research B*, 16:447–483.
- Frank, M. and P. Wolfe (1956) "An algorithm for quadratic programming," *Naval Research Logistics Quarterly*, 3:95–110.
- Friesz, T.L. and P.T. Harker (1985) "Freight network equilibrium: a review of the state of the art," in: A.F. Aughety, ed., *Analytical studies in transportation economics*. Cambridge: Cambridge University Press.
- Kinderlehrer, D. and G. Stampacchia (1980) *An introduction to variational inequalities and their applications*. New York: Academic Press.
- Knight, F.H. (1924) "Some fallacies in the interpretation of social cost," *Quarterly Journal of Economics*, 38:582–606.
- Kolata, G. (1990) "What if they closed 42nd Street and nobody noticed?" *The New York Times*, Dec. 25:C1.
- Korilis, Y.A., A.A. Lazar and A. Orda (1999) "Avoiding the Braess paradox in non-cooperative networks," *Journal of Applied Probability*, 36:211–222.

- Leurent, F. (1998) "Multicriteria assignment modeling: making explicit the determinants of mode and path choice," in: P. Marcotte and S. Nguyen, eds, *Equilibrium and advanced transportation modelling*. Boston: Kluwer.
- Marcotte, P. and S. Nguyen, eds (1998) *Equilibrium and advanced transportation modelling*. Boston: Kluwer.
- Nagurney, A. (1999) *Network economics: a variational inequality approach*, 2nd edn. Dordrecht: Kluwer.
- Nagurney, A. (2000) *Sustainable transportation networks*. Cheltenham: Elgar.
- Nagurney, A. and J. Dong (2002a) *Supernetworks: decision-making for the information age*. Cheltenham: Edward Elgar.
- Nagurney, A. and J. Dong (2002b) "A multiclass, multicriteria network equilibrium model with elastic demand," *Transportation Research B*, 36:445–469.
- Nagurney, A. and J. Dong (2002c) "Urban location and transportation in the information age: a multiclass, multicriteria network equilibrium perspective," *Environment and Planning B*, 29:53–74.
- Nagurney, A. and D. Zhang (1996) *Projected dynamical systems and variational inequalities with applications*. Boston: Kluwer.
- Nagurney, A., J. Dong and P.L. Mokhtarian (2002a) "Teleshopping versus shopping: a multicriteria network equilibrium framework," *Mathematical and Computer Modelling*, 34:783–798.
- Nagurney, A., J. Dong and P.L. Mokhtarian (2002b) "Multicriteria network equilibrium modeling with variable weights for decision-making in the information age with applications to telecommuting and teleshopping," *Journal of Economic Dynamics and Control*, 26:1629–1650.
- Pigou, A.C. (1920) *The economics of welfare*. London: Macmillan.
- Quandt, R.E. (1967) "A probabilistic abstract mode model," in: *Studies in travel demand VIII*. Princeton: Mathematica.
- Ran, B. and D.E. Boyce (1996) *Modeling dynamic transportation networks*. Berlin: Springer-Verlag.
- Samuelson, P.A. (1952) "Spatial price equilibrium and linear programming," *American Economic Review*, 42:283–303.
- Schneider, M. (1968) "Access and land development," in: *Urban development models*, Special Report 97. Washington, DC: Highway Research Board.
- Sheffi, Y. (1985) *Urban transportation networks: equilibrium analysis with mathematical programming methods*. Englewood Cliffs: Prentice-Hall.
- Taylor, M.A.P., ed. (2002) *Transportation and traffic theory in the 21st century*. Amsterdam: Pergamon.
- Volmuller, J. and R. Hamerslag, eds (1984) *Proceedings of the Ninth International Symposium on Transportation and Traffic Theory*. Utrecht: VNU Press.
- Wardrop, J.G. (1952) "Some theoretical aspects of road traffic research," *Proceedings of the Institute of Civil Engineers*, Part II:325–378.
- Wu, J.H., M. Florian and S.G. He (2000) *EMME/2 implementation of the SCAG-II model: data structure, system analysis and computation*, INRO Solutions Internal Report. Quebec: Southern California Association of Governments Montreal.
- Zhang, D. and A. Nagurney (1995) "On the stability of projected dynamical systems," *Journal of Optimization Theory and Applications*, 85:97–124.

*Chapter 34*

## TRAFFIC ASSIGNMENT METHODS

WILLIAM H.K. LAM

*The Hong Kong Polytechnic University, Kowloon*

HONG K. LO

*The Hong Kong University of Science and Technology, Kowloon*

### 1. Introduction

Traffic assignment plays a key role in the traditional transportation planning procedure, which includes these four steps: trip generation, trip distribution, modal split, and traffic assignment. The purpose of traffic assignment is to distribute or load traffic on to the network according to a certain route choice principle. Over the years, refinements have been introduced to each of these four steps. Boyce (2002) provides some historical perspectives on this procedure and points to possible improvement directions. Nonetheless, this four-step procedure forms the platform for many existing transportation planning models.

Historically, in the absence of exact solution methods, heuristics were developed to produce reasonable answers to the traffic assignment problem. One such example is the land use transport optimization (LUTO) model developed in the early 1980s for strategic land use and urban transportation planning in Hong Kong. This approach was first developed in the UK (Hall et al., 1980; Willumsen et al., 1993), and was subsequently refined and deployed in Hong Kong (Choi, 1986). LUTO developed a stochastic multi-path network loading heuristic for the traffic assignment problem.

There are other heuristic procedures for traffic assignment developed in the past. One approach assigns traffic to the shortest route between each origin-destination (O-D) pair in an all-or-nothing manner, without considering congestion effect or link capacity. Some other approaches assign traffic by taking into account the congestion effect of traffic load in the last iteration. There are also heuristics that divide the traffic load into portions and load them into the network incrementally, as congestion is gradually building up. These heuristic procedures, though somewhat intuitive, provide no guarantee that a certain route choice principle is satisfied (Sheffi, 1985).

In general, it is important to distinguish among three aspects of the traffic assignment problem. The first is to develop sound principles to model travelers' route choice behaviors. The second is to formulate these route choice principles mathematically. Finally, one must find a way to solve the formulations efficiently.

In this chapter, we provide an overview of route choice principles proposed in the past, followed by an in-depth analysis of three types of traffic assignment models. We will discuss their formulations, calibration, and validation through a case study. Specifically, this chapter is organized as follows. First, we review a few route choice principles. Secondly, the user equilibrium (UE) and stochastic user equilibrium (SUE) models are depicted, together with a brief discussion on the logit-based SUE model. Thirdly, the formulation of the probit-based SUE model (Lam and Chan, 1998a) is depicted to facilitate the discussion on model calibration. Fourthly, a case study in Hong Kong is presented to compare the various traffic assignment models. Finally, some concluding comments are given.

## 2. Route choice principles

The earliest – and subsequently widely adopted – route choice principle was proposed by Wardrop (1952). Known as the UE principle, it states that for each O-D pair, at UE, the travel times on all used routes are equal, and less than or equal to the travel times of any unused routes. This principle is behaviorally appealing, as it describes travelers' general tendency of selecting the shortest routes. Moreover, when UE is achieved, users have no incentive to switch to other routes, so that the network traffic pattern achieves equilibrium.

Other route choice principles have been proposed over the years. Among them, the principle of system optimal (SO) traffic assignment is an important one. This refers to the traffic assignment pattern that achieves the minimum total network travel time. However, the SO pattern is unlikely to occur by itself, as the UE condition is not satisfied, meaning that some travelers can lower their travel times by switching to other routes. Nevertheless, the SO result provides a benchmark of what the minimum network travel time could be. It is also a target traffic pattern that congestion pricing aims to achieve (Yang and Lam, 1996). By combining tolling to influence travelers' route choice considerations, the SO travel pattern can be achieved. Specifically, one example is to set the link toll to be equal to the marginal travel time of each link. A theoretical development of this contention can be found in Sheffi (1985).

Another important route choice principle developed in the 1970s and 1980s is the introduction of stochastic considerations; the contention is that few travelers have perfect knowledge about the network conditions. Therefore, UE is only an idealization, wherein travelers are assumed to have perfect information on the network conditions. Within this stochastic framework, travelers are modeled to

minimize their perceived (rather than actual) travel times in their route choice considerations. At SUE, no travelers can improve their perceived travel times by unilaterally changing routes. Based on discrete-choice modeling techniques, a number of stochastic traffic assignment models have been developed. Some of the earliest examples can be found in Daganzo and Sheffi (1977). The essence of stochastic traffic assignment is to model the variations or errors of travelers' perceived travel times. Different error distributions lead to different stochastic traffic assignment models. By far the most commonly adopted ones include the logit and probit models. The logit model considers the error as a Gumbel distribution, whereas the probit model considers the error as a normal distribution.

More recently, traffic assignment has been investigated by modeling traffic, route choice, and the network in a dynamic manner, which matches reality more closely. This approach is generally referred to as dynamic traffic assignment (DTA). Some examples of the DTA approach include work by Lam and Huang (1995), Ran and Boyce (1996), Lo (1999), and Lo and Szeto (2002). DTA models can provide better representations of traffic and route choice behaviors, and can capture time-dependent or dynamic network capacity changes due to incidents, traffic signals, etc. – but this is achieved at the expense of a significant increase in complexity. DTA is still in its infancy and in need of refinement (Lam and Huang, 2002).

Another new direction is to relax the assumption that the network will always operate at its ideal condition. This approach came to the fore in the mid-1990s, after we witnessed the impact of major earthquakes around the world on transportation networks. Interest in this approach also stems from heightened awareness of the security following the September 11 terrorist attack in the USA, the vulnerability of transportation networks has become an important issue. On a lesser scale, disruption of transportation networks occurs daily due to incidents or accidents. There is thus a need to study travelers' route choice considerations and network performance for the case of degradable or stochastic networks. Some early opinions on this approach can be found in Bell and Cassir (2000), Lo (2002), and Lo and Tung (2003). We expect that more will be accomplished in the future to produce more definitive results in this area of research.

In this section we have provided a brief synopsis of the main considerations in traffic assignment modeling, including deterministic/stochastic and static/dynamic route choice principles in deterministic/stochastic and static/dynamic networks. In this taxonomy, not all the approaches can be combined. For example it is not meaningful to apply a dynamic route choice principle to a static network. On the other hand, not all the meaningful combinations are well studied: as discussed earlier, DTA is still on the frontiers of research, as is the application of the dynamic stochastic route choice principle on a dynamic stochastic network.

In the remainder of this chapter we will focus on the development and results of deterministic and stochastic route choice principles on static networks.

### 3. Three traffic assignment models

#### 3.1. Deterministic UE model

Wardrop (1952) defined the route choice principle developed by him as follows: “For each O-D pair, at user equilibrium, the travel times on all used paths are equal, and (also) less than or equal to the travel time that would be experienced by a single vehicle on any unused path” – this is also known as the UE principle. Mathematically, this principle can be expressed by the non-linear complementarity conditions

$$\begin{aligned} h_k(c_k - \pi^{rs}) &= 0, \\ c_k - \pi^{rs} &\geq 0, \\ h_k &\geq 0, \end{aligned} \tag{1}$$

where  $h_k$  is the traffic flow on route  $k$  between O-D pair  $rs$ ,  $c_k$  is its route travel time, and  $\pi^{rs}$  refers to the shortest travel time between O-D pair  $rs$ . If route  $k$  is used ( $h_k > 0$ ), then  $c_k - \pi^{rs} = 0$  or  $c_k = \pi^{rs}$ . On the other hand, if route  $k$  is not used ( $h_k = 0$ ) there is no restriction on  $c_k$  other than it is greater than or equal to  $\pi^{rs}$ . Therefore, this set of conditions precisely describes the UE principle. In other words, the UE definition means that at equilibrium the routes connecting each O-D pair can be divided into two groups. The first group includes routes that carry traffic flows; the travel times on all these routes are the same. The other group includes routes that do not carry any flow; the travel time on each of these routes is at least as long as the travel time on the routes of the first group.

Together with the demand conservation constraint

$$\sum_k h_k = q^{rs}, \tag{2}$$

where  $q^{rs}$  is the demand between O-D pair  $rs$ . Conditions (1) and (2) can be expressed formally as a non-linear complementarity problem (NCP) or converted to an equivalent variational inequality problem (VIP), which provides a general formulation for the UE route choice principle.

As such, the relationship between link flow, path flow, link choice proportion and O-D demand can be defined as follows:

$$v_l = \sum_{rsk} h_k \delta_{lk}^{rs}, \tag{3}$$

$$P_l^{rs} = \sum_k (h_k / q^{rs}) \delta_{lk}^{rs}, \tag{4}$$

where  $v_l$  is the flow on link  $l$  and  $P_l^{rs}$  is the choice proportion on link  $l$ .  $\delta_{lk}^{rs} = 1$  if link  $l$  is a part of route  $k$  between O-D pair  $rs$ , and  $\delta_{lk}^{rs} = 0$  otherwise.

The link travel time functions can be expressed as the vector  $\mathbf{T} = \{\dots, t_l, \dots\}$ , where  $t_l(v_l)$  is the travel time function of link  $l$  and the link flows as  $\mathbf{v} = \{\dots, v_l, \dots\}$ . If the diagonal elements of the Jacobian matrix  $\nabla \mathbf{T}_v$  are non-negative, and all the off-diagonal elements are zero, then expressions (1) and (2) can be shown to be equivalent to the first-order conditions of the following mathematical program (MP):

$$\begin{aligned} & \min_{v_l} \sum_l \int_0^{v_l} t_l(x) dx, \\ & \sum_k h_k = q^r, \\ & h_k \geq 0. \end{aligned} \tag{5}$$

Thus, one may either solve the NCP or VIP as expressed in conditions (1) and (2), or solve MP (5) if the condition on the link travel time functions hold – i.e. the travel time on each link depends only on its own link flow but not on flows of other links. The solution algorithms to solve the NCP/VIP or MP formulations are well established. For the former, see Nagurney (1999); for the latter, see Sheffi (1985).

### 3.2. Logit-based SUE model

The UE traffic flow pattern arises if each traveler has perfect knowledge about the network conditions and all travelers have identical perceptions of travel times, which is an idealization. In reality, travelers may have certain perception variations or errors of the actual travel times. Travelers, therefore, select routes to improve their perceived travel times, rather than their actual travel times, which are not perfectly known to them. The route choice principle for this case with perception variations, often called SUE, is as follows: for each O-D pair, at SUE, the perceived travel times on all used paths are equal, and (also) less than or equal to the perceived travel time on any unused path. When a system is at SUE, travelers cannot improve their perceived travel times by unilaterally changing routes. One important aspect to note is that, at SUE, the measured or actual travel times on all used paths are not equal.

The relationship between the perceived route travel time  $C_k$  and the measured route travel time  $c_k$  for route  $k$  can be expressed as

$$C_k = c_k + \xi_k, \tag{6}$$

where  $\xi_k$  is the random perceived route travel time error. Different distributions assumed for this random error component lead to different SUE models. For the extreme case, as the error becomes zero, according to eq. (6), the perceived travel time becomes the actual travel time. The SUE formulation then degenerates into

a deterministic formulation. Therefore, one may consider the UE conditions as a particular case of the SUE conditions.

In SUE formulations, one important approach is the logit-based model, where the error component follows the Gumbel distribution (Dial, 1971). According to the logit model, the route flows can be expressed as

$$h_k = \frac{\exp(-\alpha c_k)}{\sum_j \exp(-\alpha c_j)} q^{rs} = w_k q^{rs}, \quad (7)$$

where  $\alpha$  is the dispersion parameter, sometimes known as the coefficient of perception variation,  $c_j$  is the actual travel time on route  $j$  between O-D pair  $rs$ , and  $w_k$  is a variable introduced to simplify notations. One can express the logit-based SUE route choice principle as

$$\begin{aligned} h_k(h_k - w_k q^{rs}) &= 0, \\ h_k - w_k q^{rs}, \\ h_k &\geq 0. \end{aligned} \quad (8)$$

According to expressions (8), if  $h_k > 0$ , then  $h_k$  is apportioned according to the logit split expression (7). If  $h_k = 0$ , the term  $(h_k - w_k q^{rs})$  can take any value. However, this will not happen, as the logit split expression assigns a positive flow to each of the routes. Thus, through this set of non-linear complementarity conditions, route flows are all assigned according to the logit split expression (7). Similarly to the deterministic UE formulation, expressions (8) can be expressed as an NCP or VIP, and solved accordingly. Moreover, if the Jacobian matrix of the link travel time functions follows the same conditions as in the UE case discussed above, the logit-based SUE formulation can be shown to be equivalent to the following MP, as proposed in Fisk (1980):

$$\min \sum_l \int_0^{v_l} t(x) dx + \frac{1}{\alpha} \sum_k h_k (\ln h_k - 1). \quad (9)$$

In logit-based SUE formulations,  $\alpha$  is inversely proportional to the size of the perception error. When  $\alpha$  is large, most travelers will select the route with the minimum travel time. Conversely, a small  $\alpha$  implies a large perception error, causing travelers to choose routes with large actual travel times. It is thus important to validate the logit-based SUE model in real situations and to calibrate the value of  $\alpha$ . An example of this calibration for the urban road network of Salerno, Italy, can be found in Cascetta et al. (1997).

In addition to solving expressions (8) or (9) generically, the logit-based SUE formulation can be solved by two other approaches: with and without explicit path enumeration. An example of the former was developed by Cascetta et al. (1996), in which the path enumeration procedure was accomplished by the  $k$ -shortest path

method (Ben Akiva et al., 1984; De La Barra et al., 1993) or the column generation approach (Bell et al., 1993). An early example of the latter was the STOCH algorithm developed by Dial (1971). Subsequent improvements to Dial's method include the node-based algorithm (Bell, 1995), and elimination of the cycle problem in Bell's method by Lam et al. (1996). Recently, Dial (2004) has provided a theoretical review on the STOCH algorithm and added four new algorithms for acyclic networks. The development of logit-based SUE formulation is thus still an active field.

Despite these developments, some concerns have been raised about the use of logit models because they assume the axiom of the independence of irrelevant alternatives (IIA). The IIA can be stated as "where any two alternatives have a non-zero probability of being chosen, the ratio of one probability over the other is unaffected by the presence or absence of any additional alternative in the choice set" (Luce and Suppes, 1965). Since routes usually share common links among themselves or have correlated route travel times, the assumption of independent route alternatives generally does not hold. As a result of this violation, the flows on overlapping routes tend to be overestimated. In response, Cascetta et al. (1996) proposed a C-logit SUE model to overcome the IIA problem in logit-based route-choice models. The idea is to deal with similarities among overlapping routes through the travel time attribute named "the commonality factor in the utility function." The commonality factor ( $CF_k$ ) can be defined as

$$CF_k = \beta_0 \ln \sum_j \left( \frac{L_{jk}}{L_j^{1/2} L_k^{1/2}} \right)^\gamma, \quad (10)$$

where  $L_{jk}$  is the length of links common to routes  $j$  and  $k$ , and  $L_j$  and  $L_k$  are the overall lengths of routes  $j$  and  $k$ , respectively;  $\beta_0$  and  $\gamma$  are calibrated parameters. The commonality factor of a route will be large if the same O-D pair is connected by many shared links belonging to different routes. Therefore, the commonality factor can be viewed as a penalty for routes that overlap substantially with other routes, so as to correct for the overestimation. With the incorporation of the commonality factor ( $CF_k$ ), the perceived route travel time becomes

$$C_k = c_k + CF_k + \xi_k, \quad (11)$$

where  $\xi_k$  is the perceived route travel time errors on route  $k$ . Thus, the C-logit SUE model simply replaces  $c_k$  by  $c_k + CF_k$  in eq. (7), while following the exact formulation in expressions (8).

### 3.3. Probit-based SUE model

The other important SUE model is the probit-based model, which was developed by Daganzo and Sheffi (1977). The classical probit-based model follows the same

perceived travel time definition as in eq. (6) but with the error component  $\xi_k$  assumed to follow a normal distribution. The error  $\xi_l$  for the perceived link travel time is modeled to have a zero mean and variance  $\alpha_l t_l$ , where  $\alpha_l$  is the dispersion parameter of the perceived travel time error for link  $l$  and  $t_l$  is the actual travel time of link  $l$ , expressed as

$$\begin{aligned} T_l &= t_l + \xi_l, \\ \xi_l &\sim N(0, \alpha_l t_l), \end{aligned} \quad (12)$$

where  $T_l$  and  $t_l$  are the perceived and actual link travel times, respectively. As the perceived route travel time constitutes the sum of the corresponding perceived link travel times, and that the sum of normal distributions forms another normal distribution, the perceived route travel time, and therefore the associated perceived route travel time error  $\xi_k$ , is normal distributed.

In addition to the travel time error attributable to individual perceptions, there could also be measurement errors. If these measurement errors are also taken into account for each route, the perceived route travel time becomes

$$C_k = c_k + \varepsilon_k + \xi_k, \quad (13)$$

where  $\varepsilon_k$  is the measured error of route  $k$ , and is assumed to follow a normal distribution. The measurement error  $\varepsilon_l$  of each link  $l$  along route  $k$  follows a normal distribution with a mean of zero and variance  $\beta_l t_l$ , where  $\beta_l$  is the dispersion parameter of the measurement error on link  $l$ , expressed as

$$\varepsilon_l \sim N(0, \beta_l t_l). \quad (14)$$

It should be noted that in this proposed model the perceived and measured travel time errors are considered separately. This differs from previous probit-based and logit-based SUE models, where the measured travel time errors and the measured link travel time variance are not explicitly considered. The rationale for doing so is to account for the fact that in real situations the perceived and measured errors are expected to be different. This can result in incorrect link choice proportions (see eqs (4)), leading to serious errors in estimating the O-D matrix from traffic counts. In order to account for this effect, Lo et al. (1996) discussed and proposed a model that incorporates the randomness of the link choice proportions to reduce this effect. As such, while the consideration of both the perceived and measured travel time errors is not commonly practiced, there are some merits in doing so, especially if the individual effects of perceived and measured errors are to be taken into account. The procedure for modeling perceived travel time errors is discussed first, followed by the procedure for modeling measured travel time errors.

In general, the perceived travel time error is assumed to increase with the volume/capacity ratio (or the congestion level) to describe the higher perception

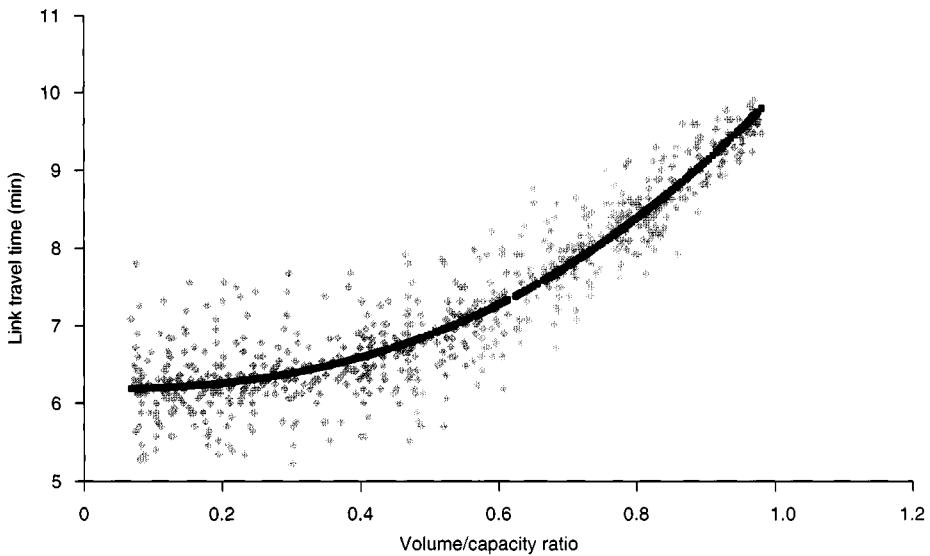


Figure 1. Link travel time versus the volume/capacity ratio.

errors associated with heavier traffic (Lam and Chan, 1998a). Mathematically, therefore, one can express the dispersion parameter of the perceived error  $\alpha_l$  as a function of the volume to capacity ratio:

$$\alpha_l = f_p \left( \frac{v_l}{s_l} \right), \quad (15)$$

where  $v_l$  and  $s_l$  are the flow and capacity of link  $l$ . One expects  $f_p(v_l/s_l)$  to be an increasing function of  $v_l/s_l$  that becomes large as  $v_l/s_l$  approaches 1. The dispersion parameter function  $f_p(v_l/s_l)$  can be calibrated using interview survey data. An interview survey can be designed to collect data on the perceived travel time and link flow in the network to establish the relationship between the perceived travel time errors and the  $v_l/s_l$  ratio.

On the other hand, the measured travel time error was found to decrease with the volume/capacity ratio (Lam and Chan, 1998a). In Figure 1, the relationship between individual travel times and their corresponding volume/capacity ratios on a certain road link in Hong Kong are plotted together with the calibrated link travel time function. As can be seen, the variation of individual travel times decreases with the volume/capacity ratio, as does the variance of the measured errors. Thus, the variance of the measured errors should be a decreasing function of the volume/capacity ratio. Hence the dispersion parameter  $\beta_l$  is expressed as

$$\beta_l = f_m \left( \frac{v_l}{s_l} \right). \quad (16)$$

Therefore, the function  $f_m(v_l/s_l)$ , which can be calibrated by conducting an appropriate survey, is a decreasing function of  $v_l/s_l$ . The change in the measured errors is small when congestion occurs. This implies that  $\beta_l$  becomes small when the ratio of  $v_l/s_l$  approaches 1.

In this probit-based SUE model, no closed-form expression exists for the route-choice proportions or route flows. Solution methods involve Monte Carlo simulation and the method of successive averages (MSA) to find the resultant link flows. Although the computation burden of simulation is substantial, this model is widely used in transportation problems (Benekohala and Zhaob, 2000; Bhat, 2000; Garrido and Mahmassani, 2000; Nielsen, 2000). Maher et al. (1997) have proposed a new heuristic solution method that does not require Monte Carlo simulation, but it may exhibit the cycle problem.

In this chapter, the simulation approach based on the simulation assignment algorithm of Powell and Sheffi (1982) is used to solve the probit-based model.

#### 4. Case study

The purpose of this case study is to:

- highlight the possible deficiency of the UE model when compared with the three SUE models;
- demonstrate the calibration of the three SUE models;
- compare the differences between the three SUE models;
- illustrate the performance of the calibrated SUE models against the observed flow data.

Figure 2 shows the network of the Tuen Mun Road Corridor, which connects the Tuen Mun and Kowloon urban areas in Hong Kong. The network consists of three zones, four nodes and 10 links.

The link travel time function with respect to link flow is

$$t_l = \phi_l + \eta_l \left( \frac{v_l}{s_l} \right)^{\rho_l}, \quad (17)$$

where  $s_l$  is the capacity of link  $l$  and  $l = 1, 2, \dots, 10$ ,  $\phi_l$  is the free flow link travel time, and  $\eta_l$  and  $\rho_l$  are the measurement parameters of link  $l$ .

The O-D matrix and the link data for the study network are given in Table 1 and Figure 3, respectively. The data from Table 1 and Figure 3 are used for calibrating the three SUE models presented in this chapter. The data are extracted from the

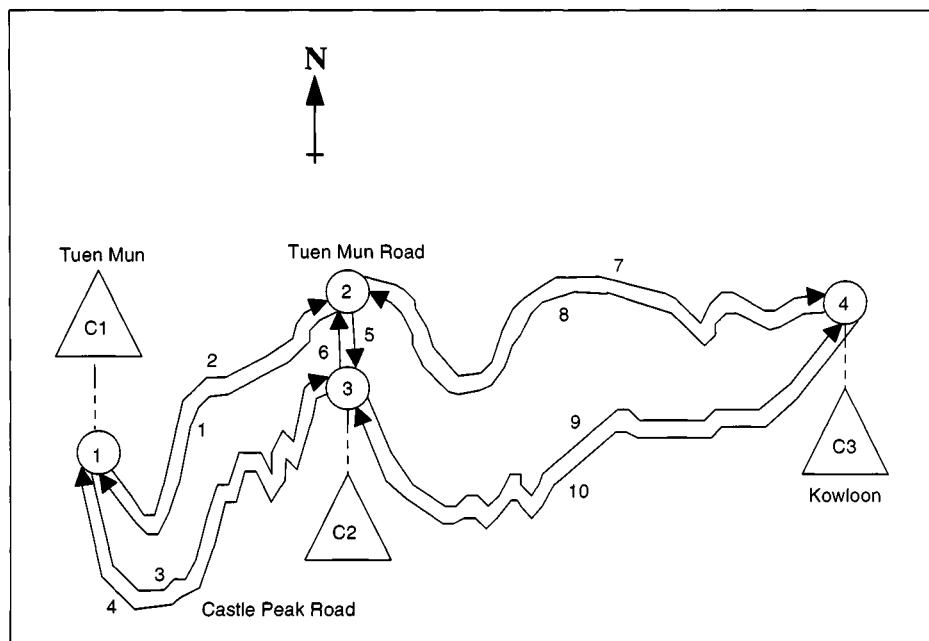


Figure 2. The Tuen Mun Road Corridor network.

Table 1  
The link data of the network

Link No.	$\phi_l$ (h)	$s_l$ (PCU/h) (a)	Parameter	
			$\rho_l$	$\eta_l$
1, 2	0.0975	5175	3.5	0.0975
3, 4	0.0922	850	3.6	0.0922
5	0.0043	730	3.6	0.0037
6	0.0043	950	3.6	0.0037
7, 8	0.0315	4800	3.6	0.0280
9,10	0.2300	1000	3.6	0.2300

Note: (a) PCU, passenger car units.

enhanced CTS-3 transport model (Hong Kong Transport Department and Wilbur Smith Associates, 1999) for the Tuen Mun Road Corridor during the evening peak period. In addition, the assignment results of the three SUE calibrated models are validated with observed link flows and compared with the results of the UE model.

Descriptions of the three SUE traffic assignment models presented in this chapter are summarized in Table 2.

As shown in Table 2, the probit-based SUE model is solved by the simulation method. The functions  $f_p(v_i/s_i)$  and  $f_m(v_i/s_i)$  used are non-linear, and were assumed to be exponential functions with proportional  $\lambda$  and  $\nu$  parameters (Lam and Chan, 1998a).

The preliminary test of the simulation shows that the change in link flows from 10 000 to 15 000 simulations is comparatively small, indicating that good convergence has been achieved. Therefore, the results of this case study are based on 10 000 simulations. The column generation approach (Bell et al., 1993) is adopted to solve the C-logit SUE model. The commonality factor (Cascetta et al., 1996) of the C-logit SUE model is as defined in eq. (10). The link-based method (Lam and Chan, 1998b) is used to solve the logit-based SUE model (Bell, 1995; Lam et al., 1996).

#### 4.1. Model calibration

To calibrate the three SUE models we use the root mean square (RMS) difference as the measure (Maher and Hughes, 1997):

$$\text{RMS} = \sqrt{\frac{1}{N} \sum_l \frac{(x_l - v_l)^2}{0.5(x_l + v_l)}}, \quad (18)$$

where  $x_l$  and  $v_l$  are respectively the observed and estimated link flows, and  $N$  is the number of links in the network.

The genetic algorithm (GA)-based technique has been used previously for calibrating traffic flow simulators (e.g. Lam and Xu, 2000), and is employed in this study to calibrate the three stochastic traffic assignment models. Based on the RMS and the initial feasible region of the coefficients, the GA-based approach

		Destination zones			
		C1	C2	C3	Total
Origin zones	C1		32	3859	3891
	C2	16		205	221
	C3	4012	309		4321
	Total	4028	341	4064	8433

Figure 3. O-D matrix (PCU/h).

Table 2  
The three SUE models

SUE model	Type of SUE model	Approach		
		Path enumeration	Commonality factor	Method
Probit based	Probit	With	NA	Simulation
C-logit	Logit	With	Yes	Column generation
Logit based	Logit	Without	No	Link based

uses an iterative approach to search for a set of better coefficients. The initial feasible region of the coefficients is defined to be between 0 and 20. This feasible region is determined on the basis of preliminary testing results. The calibration criterion is

$$\frac{\sum_l |x_l - v_l| / x_l}{N} < 0.1. \quad (19)$$

Table 3 shows the calibrated coefficients and link flows resulting from the three models (i.e. UE, probit based, C-logit, and logit-based SUE), together with the comparison of the observed link flows.

From Table 3, the estimated link flows obtained from the UE model are quite different from the observed link flows. In particular, the UE results for links 9 and 10 do not carry any flow as they are not part of the shortest route. In reality, some drivers do use links 9 and 10. In comparison, the results of the SUE methods are found to be closer to the observed link flows (even for flows in links 9 and 10).

Model validation and comparison between the various models are discussed next.

#### 4.2. Model validation

For model validation, the differences between the observed and estimated link flows are used to calculate the RMS and average relative error (ARE), and are used to assess the performance of the different traffic assignment models (Maher and Hughes, 1997). A model with a smaller RMS and smaller absolute ARE is superior to one with a larger RMS/ARE value. The ARE is defined as

$$ARE = \frac{1}{N} \sum_l \frac{|x_l - v_l|}{x_l}. \quad (20)$$

Table 3  
Link flows (PCU/h) estimated by various traffic assignment models

Link	UE	Probit-based SUE ( $\lambda = 0.86$ h, $\nu = 0.87$ h)	C-logit SUE ( $\beta_0 = 0.21$ , $\gamma = 1$ )	Logit-based SUE (dispersion parameter = $12.6 \text{ h}^{-1}$ )	Observed link flows
1	3415 (16.55)	3053 (4.20)	3045 (3.92)	3071 (4.81)	2930
2	3254 (16.21)	2855 (1.69)	2898 (3.50)	2905 (3.75)	2800
3	637 (-41.61)	1053 (-3.48)	1021 (-6.42)	1001 (-8.25)	1091
4	613 (-44.17)	992 (-9.65)	1011 (-7.92)	973 (-11.38)	1098
5	906 (-27.69)	1301 (3.83)	1301 (3.83)	1298 (3.59)	1253
6	810 (-28.00)	1249 (11.02)	1250 (11.11)	1278 (13.60)	1125
7	4064 (8.20)	3819 (1.68)	3745 (-0.29)	3785 (0.77)	3756
8	4321 (7.68)	3988 (-0.62)	3943 (-1.74)	3969 (-1.10)	4013
9	0 (-100.00)	339 (10.06)	321 (4.22)	281 (-8.77)	308
10	0 (-100.00)	347 (12.66)	380 (23.38)	353 (14.64)	308
RMS	14.66	2.04	2.27	2.52	-
ARE	-29.28	2.16	3.17	3.36	-

Note: the value in the parentheses = (resulting link flow - observed link flow)/observed link flow  $\times 100\%$ .

The RMS and ARE of the observed and estimated link flows are computed for the UE method as well as for each of the three SUE models, as presented in Table 3. It can be observed that both the RMS and ARE of the probit-based SUE model are consistently lower than those of the other three models (RMS of 14.66 for UE versus 2.04–2.52 for SUE; ARE of -29.28 for UE versus 2.16–3.36 for SUE), whereas those of the UE model are much higher, confirming the earlier observation on its deficiency in describing the observed flow data.

In Table 3, the percentage differences between the estimated and observed link flows are also calculated and shown in parentheses, a model with a smaller absolute value is superior to one with a higher percentage ratio. It is worth noting that the largest percentage difference among the individual estimates are, respectively, -100, 12.66, 23.38, and 14.64 for the UE, probit-based, C-logit, and logit-based SUE models. The UE model also performs the worse.

Based on the above results in this case study, it was found that the probit-based traffic assignment model does the best in estimating the observed link flows among the three SUE models tested, while the UE model performs the worst.

## 5. Concluding comments

In this chapter, three stochastic traffic assignment models have been presented together with their model calibration and validation results. The simulation method was adopted to solve the probit-based SUE model. The commonality factor (Cascetta et al., 1996) was used in the C-logit SUE model. The link-based method (Lam and Chan, 1998b) was used to solve the logit-based SUE model. In principle, the logit model has a simple covariance matrix, but the probit model has a more general covariance matrix structure, and is appropriate particularly for situations wherein the alternatives are not independent.

The case study results showed that the probit-based SUE model outperformed the C-logit model, which in turn slightly outperformed the standard logit model. All the SUE models substantially outperformed the UE model; as the UE model offers no parameter to calibrate its performance, this outcome is not surprising. The good performance of the probit-based SUE model is inherent from its greater flexibility in capturing the interactions between alternatives. Although the C-logit model is not as flexible as the probit-based model, it offers additional parameters to avoid the IIA violation. In this regard, it is an advancement over the standard logit-based model, and the results reflected this.

Finally, it should be noted that the computing times of the stochastic models are generally longer than that for the UE model. However, convergence of the stochastic models can usually be achieved in fewer iterations if the number of paths per O-D pair is given and restricted to eight (Cascetta et al., 1997). Recently, efficient algorithms for solving the probit model (Clark and Watling, 2002) and quasi-simulation approach (Bhat, 2000) have been developed. With advances in computer technology and improvements in solution algorithms, it is expected that applications of SUE models to large-scale networks will become common in the near future.

## Acknowledgment

This study is mainly supported by two grants from the Research Grants Council of the Hong Kong Special Administration Region awarded to the Hong Kong Polytechnic University (Project Nos. PolyU 5046/00E and N\_PolyU 515/01).

## References

- Bell, M.G.H. (1995) "Alternatives to Dial's logit assignment algorithm," *Transportation Research B*, 29:287–295.
- Bell, M.G.H. and C. Cassir (2000) "Risk averseness in user equilibrium traffic assignment: an application of game theory," in: *Proceedings of the 2nd International Conference on Traffic and Transportation Studies*. Beijing.
- Bell, M.G.H., W.H.K. Lam, G. Ploss and D. Inaudi (1993) "Stochastic user equilibrium assignment and iterative balancing," in: G.F. Newell and C.F. Daganzo, eds, *Transportation and traffic theory*. Amsterdam: Elsevier.
- Ben-Akiva, M., M.J. Bergman, A.J. Daly and R. Ramaswamy (1984) "Modelling inter urban route choice behaviour," in: I. Volmuller and R. Hamerslag, eds, *9th International Symposium on Transportation and Traffic Theory*. Utrecht: VNU Press.
- Benekohala, R.F. and W. Zhaob (2000) "Delay-based passenger car equivalents for trucks at signalized intersections," *Transportation Research A*, 34:437–457.
- Bhat, C.R. (2000) "A multi-level cross-classified model for discrete response variables," *Transportation Research B*, 34:567–582.
- Boyce, D. (2002) "Is the sequential travel forecasting paradigm counterproductive?" *ASCE Journal of Urban Planning and Development*, 128:169–183.
- Cascetta, E., A. Nuzzolo, F. Russo and A. Vitetta (1996) "A new route choice logit model overcoming IIA problems: specification and some calibration results for interurban networks," in: J.B. Lesort ed., *Transportation and traffic theory*. Amsterdam: Elsevier.
- Cascetta, E., F. Russo and A. Vitetta (1997) "Stochastic user equilibrium assignment with explicit path enumeration: comparison of models and algorithm," in: *Proceedings of the 8th IFAC Symposium on Transportation Systems*. Chania.
- Clark, S.D. and D.P. Watling (2002) "Sensitivity analysis of the probit-based stochastic user equilibrium assignment problem," *Transportation Research B*, 36:617–635.
- Choi, Y.L. (1986) "Land use transport optimization (LUTO) model for strategic planning in Hong Kong," *Asian Geography*, 5:155–176.
- Daganzo, C.F. and Y. Sheffi (1977) "On stochastic models of traffic assignment," *Transportation Science*, 11:253–274.
- De La Barra, T., B. Perez and J. Anez (1993) "Multidimensional path search and assignment," in: *Proceedings of the 21 PTRC Summer Meeting*. London.
- Dial, R.B. (1971) "A probabilistic multi-path traffic assignment model which obviates the need for path enumeration," *Transportation Research*, 5:83–111.
- Dial, R.B. (2004) "Equilibrium logit traffic assignment: elementary theory and algorithms," *Transportation Research B* (in press).
- Fisk, C. (1980) "Some developments in equilibrium traffic assignment," *Transportation Research B*, 14:243–255.
- Garrido, R.A. and H.S. Mahmassani (2000) "Forecasting freight transportation demand with the space-time multinomial probit model," *Transportation Research B*, 34:403–418.
- Hall, M.D., D. Van Vliet and L.G. Willumsen (1980) "SATURN – a simulation assignment model for the evaluation of traffic management scheme," *Traffic Engineering and Control*, 21:168–176.
- Hong Kong Transport Department and Wilbur Smith Associates (1999) *Third comprehensive transport study (CTS-3) – final report*. Kowloon: Government of Hong Kong Special Administration Region.
- Lam, W.H.K. and K.S. Chan (1998a) "A probit traffic assignment model for estimating the variance of the link flow," *8th World Conference on Transport Research*, Paper. Brussels.
- Lam, W.H.K. and K.S. Chan (1998b) "A link-based alternative to Bell's logit assignment algorithm," *HKIE Transactions*, 5:11–18.
- Lam, W.H.K. and H.J. Huang (1995) "Dynamic user optimal traffic assignment model for many to one travel demand," *Transportation Research B*, 29:243–259.
- Lam, W.H.K. and H.J. Huang (2002) "A combined activity/travel choice model for congested road networks with queues," *Transportation*, 29:5–29.
- Lam, W.H.K. and G. Xu (2000) "Calibration of traffic flow simulator for network reliability assessment," in: M.G.H. Bell and C. Cassir, eds, *Reliability of transport networks*. Baldock: Research Studies Press.

- Lam, W.H.K., K.S. Chan and M.G.H. Bell (1996) "The treatment of cycles in stochastic user equilibrium assignment," in: *4th Meeting of the EURO Working Group on Transportation*, Paper. Newcastle upon Tyne.
- Lo, H. (1999) "A dynamic traffic assignment formulation that encapsulates the cell transmission model," in: A. Cedar, ed., *Transportation and Traffic Theory*. Amsterdam: Elsevier.
- Lo, H. (2002) "Trip travel time reliability in degradable transport networks," in: M.A.P. Taylor, ed., *Transportation and traffic theory*. Amsterdam: Elsevier.
- Lo, H. and Y.W. Szeto (2002) "A cell-based variational inequality formulation of the dynamic user optimal assignment problem," *Transportation Research B*, 36:421–443.
- Lo, H. and Y.K. Tung (2003) "Network with degradable links: capacity analysis and design," *Transportation Research B*, 37:345–363.
- Lo, H.P., N. Zhang and W.H.K. Lam (1996) "Estimation of an origin-destination matrix with random link choice proportions: a statistical approach," *Transportation Research B*, 30:309–324.
- Luce, R.D. and P. Suppes (1965) "Preference, utility and subjective probability," in: R.D. Luce, R.R. Bush and E. Galanter, eds, *Handbook of mathematical psychology*. New York Wiley.
- Maher, M.J. and P.C. Hughes (1997) "A probit-based stochastic user equilibrium assignment model," *Transportation Research B*, 31:341–355.
- Nagurney, A. (1999) *Network economics: a variational inequality approach*. Boston: Kluwer.
- Nielsen, A. (2000) "A stochastic transit assignment model considering differences in passengers utility functions," *Transportation Research B*, 34:377–402.
- Powell, W.B. and Y. Sheffi (1982) "The convergence of equilibrium algorithms with predetermined step sizes," *Transportation Science*, 16:45–55.
- Ran, B. and D.E. Boyce (1996) "A link-based variational inequality formulation of ideal dynamic user-optimal route choice problem," *Transportation Research C*, 4:1–12.
- Sheffi, Y. (1985) *Urban transportation networks*. Englewood Cliff: Prentice Hall.
- Wardrop, J.G. (1952) "Some theoretical aspects of road traffic research," *Proceedings of the Institution of Civil Engineers*, II:325–378.
- Willumsen, L.G., J. Bolland, M.D. Hall and Y. Arezki (1993) "Multi-modal modelling in congested networks: SATURN and SATCHMO," *Traffic Engineering and Control*, 34:294–301.
- Yang, H. and W.H.K. Lam (1996) "Optimal road tolls under conditions of queueing and congestion," *Transportation Research A*, 30:319–332.

***Part 10***

**TIME USE**

## TIME USE AND ACTIVITY SYSTEMS

ANDREW S. HARVEY

*Saint Mary's University, Halifax*

### 1. Introduction

Time use studies clearly document that human activities do not occur in a vacuum. Each individual activity is part of a spatial, temporal socio-economic system. Each activity is part of a system of activities that integrates and facilitates ongoing day-to-day behavior and it is inextricably linked to other activities, past, present, and future. Cooperatively and/or independently, individuals and groups interact and make opportunity-changing choices impacting, at various levels, the activity systems of which they are a part. Unfortunately, the four-step approach, which for so long dominated transportation planning, failed to recognize this reality. A mechanistic approach that ignored the spatial, temporal, and individual interdependencies among transportation, land use, and population, it has left a legacy of urban areas with seriously inappropriate land use and transportation systems. The aggregate approach of the method to planning failed to provide the guidance necessary to plan efficient, equitable, and sustainable land use and transportation systems. Fortunately, major shortcomings of the four-step approach are being overcome by a shift in thinking toward an activity-based planning approach. This chapter explores the development of activity-based planning and activity systems, identifies and elucidates activity-related data needs, and discusses the important role and method of time use studies in supplying such data.

### 2. Activity systems approach

A literature review quickly confirms that there is not a theory of activity systems. Rather, the concept means many things to many people. Nevertheless, the concept is useful for structuring understanding of the world around us. The various approaches share several useful dimensions, recognizing the presence of multiple organizational entities, addressing components and contexts of individual behavior, and assuming the existence of interdependencies and patterns of behavior.

The first presentation of the activity system approach from a planning perspective was put forward by Chapin, at the University of North Carolina. Chapin and his colleagues pursued the study of activity systems incorporating time use studies at various levels. Chapin saw activities as classifiable acts, and urban activity systems as the patterned ways in which individuals, households, institutions, and firms pursue their day-in and day-out affairs in urban time and space (Chapin, 1974). He saw activity patterns as behavioral tendencies of the population within household, institutional, and firm subsystems. Behavioral studies falling within the framework include studies of travel, shopping, leisure, and even drinking behavior.

Chapin envisioned behavior as based on a “motivation → choice → outcome” sequence. He stressed the choice or preference factor in individual behavior. He recognized that an activity was also dependent upon the opportunity to act. Instead of viewing the behavioral sequence entirely as a “demand” phenomenon, the consummation of an activity was dependent upon a supply consideration as well. However, Chapin’s empirical work continued to focus primarily on the “motivation → choice → outcome” sequence.

In contrast, Hägerstrand viewed the activity process in terms of three constraints (Hägerstrand, 1970). First, capability constraints limit activities and depend upon one’s biological construction and/or tools. Second, coupling constraints define where, when, and for how long the individual has to join other individuals, tools, and materials in order to produce, consume, and transact. Third, authority constraints operate in domains where things and events are under the control of a given individual or group, e.g. zoning regulations, store hours, or even a place in a queue. Each of these constraints impacted opportunities.

Cullen argued for a more elaborate range of flexibility defined by the degree of commitment to the activity and the time–space fixity of it (Cullen et al., 1972). Cullen and his colleagues identified four categories of commitment: arranged, routine, planned, and unexpected activities. They viewed activities as occurring within bounds set by both temporal–spatial constraints and by preferences or priorities. Their approach combined aspects of both the motivational-choice mechanism of Chapin and the constraint approach of Hägerstrand. Cullen and Phelps (1975) saw behavior as an “interactive” function of its social, economic, and physical context. Chapin saw time as the common denominator for linking the various subsystems of the overall activity system.

Travel behavior, indeed all activities, are subject to several temporal constraints. Foremost is that all required and desired activities must fit into some relevant time period. Hence, travel is constrained by the time and timing required for other activities. This is reflected in Javeau’s (1972) observation that the day is structured around compulsory travel associated with essential activities. Additionally, timing is important since travel is seen as a derived demand, and is hence goal directed. The object, purpose, or person to which it is directed must be

available for it to be meaningful. This is one dimension of Hägerstrand's coupling constraint. Finally, since trips often cannot be viewed in isolation, sequence is important. Traveling to and from work, and dropping off and picking up, all imply a sequence of events.

Perception, also, is deemed to play a major role in shaping urban travel behavior (Horton and Reynolds, 1971). The conceptual model proposed by Horton and Reynolds contained three spatial elements. The first, "objective spatial structure," refers to the location of the household relative to the actual locations of all potential activities and their associated objective levels of attractiveness. Second, "action space" is the collection of all urban locations about which the individual has information, and it is the subjective utility or preference he or she associates with these locations. Third, "activity space" is the subset of all urban locations with which the individual has direct contact in engaging in day-to-day activities. They examined the formation of an individual's action space as a function of socio-economic characteristics, cognitive images of the urban environment, and preferences for travel. However, they failed to recognize explicitly the role of time in structuring space, and yet it has been shown that diurnal patterns impact the social geography of the city both in terms of status and family orientation and in terms of activity patterns (Goodchild and Janelle, 1984). Given the role of time with respect to timing, duration, sequencing, and the structuring of activity, it is imperative that it be incorporated into analysis, modeling, planning, and policy.

### 3. Time use and travel behavior

Time use and/or travel researchers have been exploring individual travel behavior, at first off and on, and later to an increasing extent, beginning in the mid-1960s. Until about 1990, time use and travel researchers worked relatively independently of each other. Since the early 1990s there has been greater collaboration. It is useful briefly to examine the work of researchers in each group over that period.

#### 3.1. *The time use perspective*

For the past half century, time use researchers have, from time to time, addressed travel issues, providing ample support to the efficacy of the time diary approach. Javeau (1972) drew on time use data collected in Belgium in February and March 1966, as part of the Multinational Time-Budget Study (MTBS), to analyze the trip to work, particularly rush hour traffic. Javeau argued that the day was structured around compulsory journeys attached to essential activities, and those time diaries were ideal instruments for measuring the constraints. Javeau observed that time diary studies would benefit from the addition of spatial data and from other data

of the type typically collected in traffic studies. He theorized that there is a direct relationship between traffic density and the extent of constraint in the environment. He argued for the integration of these concepts into urban planning and that the elimination or reduction of compulsory time will come about only with changes in urban patterns (Javeau, 1972). In October and November 1965, as part of the MBTS, spatially coded data were collected in Osnabruck, Germany, to study the outdoor activity system in an urban environment (von Rosenbladt, 1972). The study noted both a great complexity in reactions to environmental variables and that people reduce out-of-home activities if they live in an area of low facility accessibility.

It was also in the mid-1960s that time use methodology was applied in a pilot investigation for the analysis of human behavior patterns in the USA. Chapin and Hightower (1966) argued that most land use and transportation routes exist as a means for the accomplishment of some desired activity. Whereas movements are the key component of transportation planning, activities are fundamental to land use planning and constitute the link between the city and the people. They believed that the study of activity patterns recurrent in time and space might ultimately yield better theoretical explanations and improved predictive tools. An initial pilot investigation using data from interviews in two census tracts of a small city was undertaken to examine problems and potentialities of activity analysis. The objective was to provide a basis for analyzing and eventually predicting changes in urban space use, daily activities, and travel of urban residents.

In the Dimensions of Metropolitan Activity (DOMA) study undertaken in Halifax, Canada, in 1971–1972, and modeled after the MTBS, activity location was explicitly captured and coded on 1 km and 0.1 km grids, the latter in built-up areas (Elliott et al., 1976). The inclusion of spatial coordinates made possible the determination of a large number of measures and equated the time budget data with travel origin and destination surveys. It became possible to identify both the distance and direction of travel and, when distance was combined with duration, the speed of travel. However, the failure of the data to show the travel route limited distance estimations and overestimated speed. (The ability to incorporate Global Positioning System (GPS) information in modern studies removes such limitations.) Data from the DOMA study provided the basis for a typology of urban behavior settings. Home/work was the strongest organizing dimension of urban space-time, followed by entertainment, shopping and education/work (Goodchild et al., 1993).

### *3.2. Travel perspective*

In the mid-1970s, travel researchers attempted to apply activity systems thinking in a travel survey, Jones and his colleagues discussed seven key features of human

behavior: activities, time–space, constraints, households, interactions, adaptation, and change (Jones et al., 1983). They also explored four areas of application and development: household activity data, computer-based demand models, policy issues, and the dynamics of travel decisions. The authors pursued their ideas with a time–activity diary administered to 204 households in Banbury, UK, in the fall of 1976. Most importantly, they noted that activities are distinguished by several dimensions, including behavior, situation, psycho-physiological state, social orientation, and psychological orientation. They concluded that understanding activities requires the capture of all significant dimensions of an activity. Further, they noted that strong interrelationships among household members, with respect to travel, were an argument for collecting household diaries.

In 1991, Stopher introduced an activity diary in the Boston area. Its introduction and design were premised on the growing conviction, as recognized earlier by Jones and his colleagues, that most travel is a derived demand, and hence it is necessary to identify and understand travel in relation to other activities. The Boston diary combined all at-home activities together while disaggregating out-of-home activities. Subsequent studies in the mid-1990s in Oregon–Southwest Washington, Raleigh–Durham, and San Francisco, as well some other locations, have collected activity detail for both in-home and out-of-home activities. However, all used limited activity classifications and a diary approach more consistent with the typical trip diary than a typical time diary. More recent travel study diaries cast in the nature of day-planners provide a format more consistent with traditional time diaries (Stopher and Wilmot, 2001). Building on this concept, an electronic version of the day-planner format has been developed with a particular emphasis on exploring activity planning, decision-making, and behavior.

The role of preferences also gained increasing attention beginning in the 1970s. Considerable attention was directed toward the measurement and analysis of stated and revealed preferences and their integration into the decision-making process. Space prevents further discussion of these advances here; however, there is a significant literature in terms of the activity systems paradigm (Koyama et al., 1998).

In view of the significant and growing role played by time use data in travel studies, the remainder of this chapter addresses issues related to the role of time use studies in activity measurement, capture, and analysis.

#### 4. Time use measurement

Time use data providing indicators of participation in, and time allocated to, activities such as travel, labour force participation, reading, television viewing,

cultural events, and other activities are captured by instruments that range from direct questions, through simple activity lists, to time diaries or even direct observation. The appropriate approach for capturing activity data depends upon the type and accuracy of the data sought, the funds available for their collection, and the strengths and weaknesses of the several collection approaches.

A full-time diary (budget) records, in sequence, all activity events/episodes (e.g. eat breakfast, drive to work), along with start/end times, engaged in by an individual over a specified period of time, most frequently a 24 h day. The time diary is an exhaustive data collection approach, capturing all activities and covering all time within the target period. Typically, time diaries allow for the continuous description of behavior in the vernacular of the respondent, thus leaving interpretation until at least the data entry stage.

Time diary studies are nearly 100 years old. However, they were infrequently undertaken until a major stimulus for such studies came with the advent of the MTBS in the mid-1960s (Szalai, 1972). Since then, and particularly in the last decade, time diary studies have been increasingly collected by official statistical offices. In January 2003, the US Bureau of Labor Statistics started the first time use study designed to be ongoing month to month and year to year.

Several observations are warranted based on time use studies (Table 1). First, in most of the studies, only one day was captured since the number of diaries equals the number of respondents. Exceptions are Norway, Sweden with 2 days per person, and the Netherlands with 2 days in 1997 and 7 days per respondent in 1990 and 1995. Trips as a percentage of all diary episodes ranged from 16.2% in Norway in 1990 to 20.9% in the Netherlands in 1997. Finally, average daily trips per person ranged from 3.8 in the Netherlands in 1997 to 5.4 in Halifax in 1971. The average 4.3 trips per day exceeds the average number of trips typically registered by trip diaries.

Time diaries place activities in context. They chronicle the sequence and duration of activities undertaken by the diary keeper over a specified period. They provide a consistency in time–activity data by leading people through the day and causing them to account, in a constrained manner, for their time. Diaries facilitate the recording of a number of contextual dimensions of each particular act. Additionally, time diaries provide considerable latitude for the collection of ancillary contextual information connected with each episode (diary entry). The MTBS collected data on what was being done (primary activity) and when (time of day), each minute over a 24 h day, guided by 10 min blocks of time. Data on situational and activity-related contextual dimensions of the activities were also captured. The study also collected information on secondary activities, which were coded with the same classification system as that used for primary activities. Hence, each diary entry would capture the following data: the activity, e.g. listening to the radio (secondary) while eating (primary); where the activity took

**Table 1**  
Diary activity and trip dimensions: selected time use surveys

Survey	Persons	Diaries	Days	Diary events	Activities per day	Trips	Trips as percentage of events	Trips per person
Halifax DOMA (1971)	2 141	2 141	1	60 607	28	11 614	19.2	5.4
Canada Pilot (1981)	2 682	2 682	1	72 987	27	13 145	18.1	4.9
Canada (1992)	8 996	8 996	1	190 327	21	38 563	20.3	4.3
Canada (1998)	10 479	10 479	1	221 105	21	40 910	18.5	3.9
The Netherlands (1990)	3 415	23 905	7	493 553	21	101 807	20.6	4.3
The Netherlands 1995)	3 227	22 589	7	560 258	25	102 349	18.3	4.5
The Netherlands (1997)	1 328	2 693	2	49 780	18	10 338	20.9	3.8
Norway (1990)	3 088	6 174	2	176 191	29	28 616	16.2	4.6
Austria (1991)	25 233	25 233	1	513 152	20	98 073	19.1	3.9
Sweden (1991)	3 630	7 187	2	219 165	30	36 684	16.7	5.3
Total	58 438	107 256		2 423 531	23	457 340	18.9	4.3

place, coded by a 10-digit generic code (home, workplace, etc.); and who else was involved (spouse, children, etc.,) (Szalai, 1972).

Situational contextual dimensions are always present: one must be doing a secondary activity (or not); one must be somewhere; one must be with someone (or alone). Additionally, each activity can be expected to have certain subjective dimensions, e.g. feelings of enjoyment, stress, etc. Situationally determined context variables should be collected for all activities. They are normally captured by diary columns parallel to the activity (episode) column. Activity-determined context variables are uniquely bound to specific activities, e.g. mode of travel, the nature of reading material, or what is being viewed on television. Hence, the context for specific activities needs only to be collected for pertinent activities defined by the objectives of the study. From the time diary perspective, trips are simply activities. They are a part of the daily activity pattern, they may be accompanied by secondary activities such as radio listening or conversation, done alone or with someone else. Their location is the mode of travel, their purpose and trip type are defined by activity and location at the origin or destination, and they integrate into a sequence of activities (trips) in a natural way to form tours or journeys.

Where time use studies are being implemented for travel research, contextual variables related to trips and other relevant activities can be used. In the case of trips, in time diaries the mode of travel is normally captured as location, whereas trip purpose can be captured from purposes at origin/destination. Other dimensions of travel, e.g. comfort or convenience, can be captured in an "other" column designed to capture the activity context for activities, where appropriate, such as what is being watched on television, read, or listened to on the radio, or the context can be captured by question sequences referring to diary episodes.

The major strength of the time diary is that it is time constrained: the reported time allocations must exhaust a defined time period, i.e. 24 h in the day or 168 h in the week. In contrast, trip diaries/travel logs focusing on trips, the most widely used approach to obtain travel data, are selective and time unconstrained. They target only specific activities, i.e. travel, for which one wants participation, duration, and/or timing information. Respondents are asked to record all trips and provide selected information relating to each trip, during some specified period, ranging from a single day to a week, month, or longer. Similar logs have been used to collect data on television viewing, shopping, housework, and other activities. Such instruments are selective rather than exhaustive. They extract, from the total range of behavior, those elements of particular current concern, independent of other activities and the underlying life context. Since researchers now accept that travel is a derived demand and cannot be meaningfully considered independently of the activity patterns of which it is a part, they have been led to seek more inclusive and precise methods of data collection. Time diaries emerge as the device of choice to capture activity information.

#### 4.1. Time use data collection methodology and instruments

Time use and travel data collection techniques and approaches have been extensively addressed elsewhere (Harvey, 1993; Ettema et al., 1996). However, it should be noted that time use data pose several methodological concerns relating to sampling, the nature of respondents, instrument design, and content and collection regimen, not normally addressed by social surveys. Foremost, both people and time are being sampled. Specific issues that must be addressed are the population, insuring representation with respect to the age of respondents and geography; sample size and sampling method; and timing, considering the time of year, number of days, and choice of diary day. Data collected must be representative of full 24 h days and all days of the week, to allow for the appropriate daily and weekly cycles of activities. A designated day, a day which was chosen by a purposeful selection method, should be used.

Diaries should be collected for either yesterday (phone or personal interview) or tomorrow (leave behind and pick-up). At least 2 days per respondent is strongly recommended. Increasing the number of days per respondent reduces important dimensions of measurement error and increases the usefulness of the data for behavioral modeling, but also increases costs and non-response. The yesterday diary, typically by phone, requires at least two calls to complete 2 days unless they are contiguous and both can be captured on the third day, since recalled events should not be more than 2 days old. The tomorrow, or diary, approach is more flexible in terms of the number of days that can be captured.

An open-interval diary is recommended. However, diaries could use fixed-interval time slots not exceeding 15 min. Recent studies in the Netherlands, Norway, Switzerland, and the UK have used 15 min blocks, while 10 min was used in Finland. An open format has also been used in Canada and the USA. Lingsom (1979) concluded that there was no substantive difference between open- and fixed-interval diaries. After testing the two methods in a pilot study in 1979, she concluded that interval diaries were the preferred alternative for the Norwegian study, based on the fact that the diaries were to be self-completed by the respondent. These issues are treated elsewhere at greater length (Harvey, 1993).

Diary surveys consist of at least two parts: a questionnaire collecting socio-demographic, context, and subjective information useful in classifying and interpreting the second part, which is the diary. Diaries collect a number of temporal and contextual variables. It has been suggested that the questionnaire content of travel studies can be divided into three non-activity groups: household, personal, and vehicle (Stopher and Jones, 2003). Household information includes items related to location, dwelling structure, family structure, and income. Personal information relates to demographic characteristics, role, occupation, and static mobility. Vehicle information relates to type, year, use, and fuel.

#### *4.2. Activities and context*

Ideally, the diary should collect data on the primary activity, secondary activity, location of the activity, with whom, and for whom. Primary and secondary activities need to be captured both to fully reflect and measure daily activity patterns and to facilitate reporting by respondents. If the secondary activity is not collected, certain crucial activity information is lost. Also, in many cases, respondents find it difficult to report only one activity in a given time slot. In addition to the activity dimensions noted above, among both travel and time use researchers there is strong and increasing interest in obtaining information on subjective dimensions of recorded activities. Travel preferences and perceptions, tension/relaxation, enjoyment, and satisfaction dimensions are useful in better understanding and modeling behavior. While one may question how all the information will be used in analysis, the detail is collected not only for analytical reasons. Additional detail can be important in aiding recall at the time of collection and in coding at the coding and data editing and entry stages.

#### *Activity reporting and coding*

The received practice is to use a free-form diary, or an interval diary measured off in 10 or 15 min intervals, allowing the respondents to describe what they were doing in their own words. Free form is preferred for interviewer-administered diaries, while a measured form is preferred for self-completion diaries. A particularly friendly diary based on a day-planner format has been developed that provides for both interval and free-form reporting (Stopher and Wilmot, 2001). A hierarchical coding scheme, sufficiently detailed to provide an unambiguous and exhaustive record of activity without undue confusion, is needed. Of the major national studies, coding runs from 32 to 200 or more activities. In particular, primary activities must add up to 1440 min per day.

#### *With whom coding*

Data should be collected reflecting social contact with at least the following categories:

- (1) alone;
- (2) spouse;
- (3) children;
- (4) other household members;
- (5) co-workers/schoolmates;
- (6) friends/relatives outside household;
- (7) others outside the household.

For whom coding a distinction must be made during capture and/or coding between “being in the presence of” and “acting with.” The data should, at a minimum, permit the researcher to identify the time that the respondent was “acting or interacting with” others. The diary and survey procedure should also provide for capturing at least two contacts, e.g. with spouse (2) and others (7).

### *Location coding*

Location data provide valuable information for use in coding the diaries and for analysis. At least the following generic locations are recommended:

- (1) home inside;
- (2) home outside;
- (3) workplace (away from home);
- (4) other person’s home;
- (5) elsewhere out of home, traveling
  - (6) on foot,
  - (7) by bicycle,
  - (8) by car,
  - (9) by public transit,
  - (10) by other/unknown;
- (11) other/unknown location.

Time use data sets containing specific geographic locations for activities exist but are rare. While the capture of precise geographic location data has been seen as an ill-affordable luxury in most time use studies, such information is imperative for transport studies. Traditionally, geographic information was only captured and/or recorded at the zone level. However, it is important to make it more precise, and GPS technology provides the means to do so.

### *For whom coding*

This was first introduced in the 1991 German national time use study. For whom an activity is done promises to be a valuable variable for understanding and modeling behavior. It is a key variable in understanding motivation. In the German study, the focus was on non-paid work activities. Respondents were asked whether the diary activity was for their own household, another household, their own and another household, or voluntary work. A French study also used a four-way breakdown encompassing paid and unpaid activities consisting of oneself and household, work, another household, and organizations. A study based in Australia has used a far more elaborate classification containing 19 categories. Many of the recent studies have used some variant of “for whom.” In some cases, however, it was only collected by questions focusing on volunteer or helping

activities. Since for whom an activity is done is a situational contextual dimension it should be captured for all activities.

### *Subjective variables*

A variety of subjective variables have been used in diaries to capture psychological dimensions such as tension, stress, enjoyment, and satisfaction. Typically, such variables are captured using a scale with values ranging between 0 and 10. Subjective variables on a diary put the situation in context rather than depending upon random recall of events and emotions. Subjective variables have also been used to capture perceptions related to alternatives to activities being undertaken. Could the activity have been done at a different time or at a different place? Such information provides insight into the time-space fixity of activities.

## **5. Time use analysis**

### *5.1. Unit of analysis*

When analyzing time diaries, the researcher has several units of analysis. This provides a richness and breadth to the analysis that is both challenging and fruitful. At a minimum, meaningful analysis of time use data would encompass several of the following.

#### *Episodes or events*

Episodes or events are the smallest unit of analysis in time diary studies. It is a line in the diary. At the episode level, it is possible to determine the context of a specific act. One would know, at a very minimum, what was done, when it started, and for how long it lasted. Depending upon the dimensions collected, one would know the presence or absence of secondary activities, where the act was done, and in whose presence (children, spouse, etc.), various objective and subjective traits, the level of tension, etc. Each dimension can be analyzed in terms of the information contained in the episode vector. When examining episodes, it is also possible to examine the sequence in which activities take place. The most interesting aspects of a diary stem not from the amount of time devoted to an activity or from the fact that it has been participated in but from the way in which a given episode of an activity fits into a sequence of activities by the respondent during the day. At the activity level, such detail is lost.

Episode-based sampling provides special analysis opportunities. Drawing on the several episode dimensions, it is possible to identify special populations for further analysis. Assume one wants to study individuals who work at home. Any individuals who did so can be identified from episodes of paid work at home. Once

such individuals have been identified, they can be studied as a group. In one study, all respondents who indicated an episode of pet care were flagged, and the data divided into two groups: from respondents with pets, and from respondents who either did not have them or did not do any pet care on their diary day. Controlling for other relevant variables, it was found that the pet owners averaged five trips more per day compared with the other respondents.

### *Activities*

Activities defined in terms of a finite set of activity categories such as travel, eating, or working are a major focus of attention. Daily trips, meals, and work episodes are combined into their activities, travel, eating, and paid work, for example. An advantage of dealing with the activity is its circumscribed nature. It is simpler to deal with 96 or 37 activities than to deal with an unknown number of episodes. The major disadvantage, of course, is that it is a summary measure. Context is essentially lost. While detail can be attached to the activities and the number of activities can be expanded to be more context-specific, considerable detail about individual behavior patterns is lost through the aggregation of episodes.

### *Day*

Day may also be the unit of analysis. Typically, however, the day has generally only been examined in terms of weekdays versus weekend days. However, since the character of the day may impose constraints or provide opportunities, there is value in understanding better the role it may play in shaping behavior. Collection of more than 1 day per respondent makes it possible to study the effect of behavior on one day on behavior on another day, or to study the recurrence of activities. Also, there is a case to be made for having at least 2 days per respondent in order to reduce variance and get more efficient estimates.

### *Participants/non-participants*

Participants ("doers" on diary day) are people engaging in a given activity, for example travelers. An analysis of participants yields insight into their behavior. How long do individual travelers spend traveling? How do travelers allocate travel time across modes, trip purposes, etc.?

### *Respondents*

Respondents are the typical unit of analysis. The researcher is concerned with whether or not an individual has participated in certain activities, how they allocate their time across activities, or how individuals apportion the day or other period among various activities.

### *Population and subpopulations*

Population estimates require sufficient sampling information to provide appropriate data weighting. However, subpopulations are generally a major focus of any analysis. What proportion of any given population group engages in given activities? How long do they spend doing them? Answers to such questions flow naturally from time use studies. For example, what proportion of the population engages in travel, paid work, etc., and how long do they spend on each?

### *5.2. Activity measures*

Minimally, activities can be measured in terms of participation rates by a binary measure that distinguishes “participation” from “non-participation.” However, typically, participation is qualified in terms of intensity along a time dimension, which may range from minutes (a day) – “duration” – to times (a day/year) – “frequency.” Since empirical studies have shown that frequency and duration measures of the same behavior are not always highly correlated, both should be collected. In fact, several measures are required to render a useful understanding of recorded behavior. It is helpful to examine briefly all such measures.

#### *Activity participation*

Activity participation, registered by stated involvement or time recorded in a diary, is involvement in an activity, independent of the amount of time devoted to it. Typically, it is presented as a participation rate (doers/population) per unit time for a given group. In carrying out a day-based activity survey, one is sampling both individuals and time (days); thus, participation reflected in time diaries depends upon two rates – individual participation and daily participation. A participation rate calculated from time diaries cannot be assumed to reflect only population participation unless the activity is done every day by those who do it. If everyone does the activity, e.g. sleep, then the measured rate does reflect frequency or the likelihood of sleeping on an average day. For example, assuming everyone travels some time during the week (100% individual participation), a travel participation rate of 0.857 (6/7) indicates a daily participation frequency averaging travel as 6 out of 7 days.

#### *Activity frequency*

Activity frequency is the number of episodes of a given activity occurring during a specified period of time, e.g. trips per day. The population daily average will be lower than it will be for travelers, in direct proportion to the rate of participation

in travel. Frequency is often collected using direct questions or activity lists; sometimes the frequency counts are used as a surrogate measure of time allocation. However, the lack of any constraint on the expressed number can generate considerable error. Frequency depends heavily upon variations in data collection methods and coding. Furthermore, as noted above, it is of limited value in comparing activities that are likely to differ significantly in the amount of time devoted to each episode. Time diaries impose a temporal constraint on reported episodes, ensuring the most accurate results.

### *Activity duration*

Activity duration is the quantity of time devoted to an activity. In traditional time allocation studies, it refers to minutes or hours per day or week. Duration is the key temporal indicator. Two duration measures are typically used. One is the population duration (total time divided by the total population). The other is the participant duration (total time divided by number of active participants). Assuming that the average travel time across the whole population, the population duration, is 60 min per person, and that on the average day the activity participation rate is 0.857%, the participant duration would be 70 min ( $60/0.857$ ). A time-exhaustive diary, covering 100% of the diary time, improves reporting accuracy since durations must add up to 24 h, so any deviation must be offset by a change elsewhere in the diary.

### *Episode duration*

Episode duration, the time allocated to a single episode (event) in an activity diary, is the prime duration measure. Activity durations mentioned above are aggregations of all similar episodes in a given diary. The activity sum masks the amount of time allocated to single events. The fact that participants travel 70 min per day says nothing of individual trips. There could be two trips of 35 min, seven trips of 10 min or 14 trips of 5 min recorded in a day or some other combination. Travel behavior will differ, depending upon actual trip lengths faced by individuals, independent of total time traveled.

### *Temporal location/timing*

Temporal location/timing in terms of time of day, week, month, or year an activity is undertaken portrays the rhythm of society, structures behavior, and is important in the analysis of both trips and other activities. The time of day that people depart for work, shop, prepare meals, eat or are at school contributes to structuring travel, traffic, and society. Clearly, peak-hour travel is a timing issue. At another level, temporal location may refer to pay day, a public holiday, or the time of year

when vacations are scheduled. Knowledge of the temporal location of activities is imperative if one wants to understand rhythm and change in a society.

### *Activity sequence*

Activity sequence differs from temporal location. Sequence relates the occurrence of a given activity episode to activity episodes that precede and follow it. It helps one to understand trip chains and other means individuals use to organize a day. Sequence also enhances our understanding of activity participation and the way travel may enhance or restrict it. Individuals may sequence activities (trip chain) going to work, then directly to shop and then to leisure, before returning home. Recent approaches to the analysis of sequence promise to give us a greater insight into the sequence dynamic in daily activity patterns.

## **6. Advantages and challenges of the time use approach**

A time use approach has a number of advantages. First, respondents often find it easier to describe their day as it unfolds rather than extract and report on certain activities out of context. Jones et al. (1983) found a major strength of their activity diary was that activity recall contrasted strongly and favorably with trip recall. Secondly, responses tend to be more complete and more accurate. Respondents had little trouble recalling activities, changes of place, and, hence, travel. Changes in one dimension, for example a change in whom an individual is with, may remind them of a change in another dimension, such as a change of activity or of travel. Thirdly, the detailed information collected facilitates diary editing and coding. Fourthly, research suggests that time diaries yield more trips than trip-only diaries do. Lastly, time diaries provide almost limitless flexibility at the analysis stage. One can look at temporal dimensions including duration, timing, sequence, as well as other contextual dimensions, all constrained by the fact that all possible activities and a fixed amount of time are accounted for. They also make possible emergent analysis of special issues.

Time diary studies are not without challenges. However, small relative to the benefits, these difficulties can be overcome to yield a high return on investment. High cost is often cited as an impediment to time diary studies in general. However, relative to travel diary studies, there should be little difference in cost. Other problems that have been identified can be overcome with proper study design and execution. Jones et al. (1985) identified problems with completion of their diary. First, they found technical problems in recording activities. With respect to location recording, they found that location was often too general. Also, travel was often included in the activity, e.g. "went shopping." Secondly, they found problems in activity definition, particularly in the way that the respondents

had recorded overlapping and co-occurring activities. Other concerns related to whom respondents were with and trip context. Respondents need to be carefully instructed on what is being sought. Careful instruction and examples can overcome the problems.

## 7. Conclusions

Time-diary-based activity studies have been central to the activity systems approach from its beginning in the middle of the last century. Their growth, in number and sophistication, has provided an abundance of data for travel analysis, and insights which can be used to design and execute surveys yielding more accurate and enriched travel data. They offer the opportunity to capture the many activity system elements found to be important in understanding and modeling travel behavior. These elements include activity context, preferences, and perceptions. Additionally, activity studies provide a wide range of measures designed to shed light on a variety of analytical units ranging from diary episodes to the population. Time diary data, coupled with geographic location data, decision-making analysis, and supply or opportunities data, offer analysts, modelers, and planners a solid working base to address urban travel and land use issues.

## References

- Chapin, F.S. (1974) *Human activity patterns in the city: things people do in time and in space*. New York: Wiley.
- Chapin, F.S. and H.C. Hightower (1966) "Household activity patterns and land use," *AIP Journal*, Aug.:222–231.
- Cullen, I., V. Godson and S. Major (1972) "The structure of activity patterns," in: A. Wilson, ed., *Patterns and processes in urban and regional systems. Papers in regional science 3*. Berlin: Springer-Verlag.
- Cullen, I.G. and E. Phelps (1975) *Diary techniques and the problems of urban life*, Grant No. HR2336. London: Social Sciences Research Council.
- Elliott, D.H., A.S. Harvey and D. Procos (1976) "An overview of the Halifax time-budget study," *Society and Leisure*, 3:145–159.
- Ettema, D.F., H.J.P. Timmermans and L.A.M. van Veghel (1996) *Effects of data collection method in travel and activity research*. Eindhoven: European Institute of Retailing and Services Studies, Eindhoven University of Technology.
- Goodchild, M.F. and D.G. Janelle (1984) "The city around the clock: space-time patterns of urban ecological structure," *Environment and Planning A*, 15:807–820.
- Goodchild, M.F., B. Klinkenberg and D.G. Janelle (1993) "A factorial model of aggregate spatio-temporal behavior: application to the diurnal cycle," *Geographical Analysis*, 25:277–294.
- Hägerstrand, T. (1970) "What about the people in regional science," *Papers and Proceedings of the Regional Science Association*, 2:7–21.
- Harvey, A.S. (1993) "Guidelines for time-use data collection," *Social Indicators Research*, 30:197–228.
- Horton, F. and D.R. Reynolds (1971) "Effects of urban spatial structure on individual behaviour," *Economic Geography*, 47:36–48.

- Javeau, C. (1972) "The trip to work: the application of the time-budget method to problems arising from commuting between residence and workplace," in: A. Szalai, ed., *The use of time: daily activities in urban and suburban populations in twelve countries*. The Hague: Mouton.
- Jones, P.M., M.C. Dix., M.I. Clarke and I.G. Heggie (1983) *Understanding travel behaviour*. Frome: Rowe.
- Koyama, N., F. Hanai, T. Yokota, D. Hensher, J. Louviere and J. Swait (1998) "Combining sources of preference data," *Journal of Econometrics*, 89:197–221.
- Lingsom, S. (1980) *Open and fixed internal time diaries: a pilot study on time use 1979*. Oslo: Statistisk Sentralbyra.
- Stopher, P.R. and P.M. Jones (2003) "Developing standards of transport survey quality," in: P.R. Stopher and P.M. Jones, eds, *Transport survey quality and innovation*. Amsterdam: Pergamon.
- Stopher, P.R. and C.G. Wilmot (2001) "Development of a prototype time-use diary and application," *Transport Research Record*, 1768:89–98.
- Szalai, A. (1972) *The use of time: daily activities of urban and suburban populations in twelve countries*. The Hague: Mouton.
- von Rosenbladt, B. (1972) "The outdoor activity system in an urban environment," in: A. Szalai, ed., *The use of time: daily activities of urban and suburban populations in twelve countries, Part II*. The Hague: Mouton.

*Chapter 36*

## ACTIVITIES IN SPACE AND TIME

HARVEY J. MILLER

*University of Utah, Salt Lake City, UT*

### 1. Introduction

Human lives consist of activities such as working, raising families, socializing, shopping, and recreation. These activities require time and space, and are often only available at particular locations for limited durations. People differ with respect to the location and timing of key activities in their lives (e.g. home, work) as well as available time and transportation and communication resources to conduct these activities.

Cities, transportation, and communications systems exist to alter space and time. A city compresses many human lives into a relatively small space to reduce time and other resources required for humans to interact and conduct shared activities. Transportation systems reduce the amount of time required for movement across space. Communication technologies eliminate space for some types of interaction. In turn, cities, transportation, and telecommunications systems shape human activities by altering the relationships between space and time in human interaction and activity participation.

The observations listed in the preceding paragraphs seem uncontroversial, perhaps even trite. Yet many transportation and urban models ignore the basic spatial and temporal conditions of human activities. Transportation and communication have critical roles in fitting people and activities together in space and time to create functioning socio-economic systems. Understanding how transportation and communication technologies facilitate and constrain the spatio-temporal web of human interaction is crucial for designing efficient, livable, and sustainable cities and economies (Hägerstrand, 1970).

This chapter discusses theories and methods for analyzing the interrelationship of human activities in space and time, focusing on time geography: an elegant and powerful framework developed by Hägerstrand and his colleagues at Lund University in Sweden during the 1950s and 1960s. The relevance of time geography for analyzing real world transportation and communication systems has improved tremendously in recent years with the rise of geographic information technologies

such as geographic information systems (GIS) and the Global Positioning System. At the same time, information technologies such as the Internet and mobile telephony have drastically altered the relationships between geo-space and time in the real world, creating fascinating and imperative research challenges. The “new time geography” has great potential for guiding land use and transportation systems toward more livable and sustainable outcomes.

## 2. Time geography

Time geography focuses on the interrelationships between activities in time and space, and the constraints imposed by these interrelationships. Rather than attempting to explain or predict an individual’s allocation of time among potential activities in space, time geography highlights the factors that restrict an individual’s choice. Although often applied to daily and weekly time frames at the urban scale, time geography can also accommodate scales as extensive as a person’s lifetime (Hägerstrand, 1970).

### 2.1. *Activities in space and time*

The fundamental tenet underlying time geography is that all activities have both spatial and temporal dimensions that cannot be meaningfully separated. The “choreography” or sequence of events that comprise a person’s existence at any temporal scale (daily, monthly, lifetime) consists of activities that have temporal duration and spatial extent. Required or desired activities such as home, work, shopping, recreation, and socializing occur only at a few locations in space, and for limited durations. A person must trade time for space through movement or communication to participate in these activities (Pred, 1977).

Activities differ with respect to their pliability in space and time. Fixed activities refer to events that are relatively difficult to reschedule or relocate. For example, people are often required to work at a specific location for a designated duration. A person’s home is usually fixed in place (at least over the short run), and maintenance or familial obligations (as well as basic biological needs such as sleep) require presence for regular intervals. Flexible activities are relatively easy to reschedule and relocate. A person can shop, recreate, or socialize at otherwise idle times between work and home hours; he or she also has a choice over where and when this occurs. There are limits on flexible activities as well (e.g. retail outlets have limited hours and few locations, and one cannot socialize if friends are not available), but the activity is flexible if there are some degrees of freedom to the individuals involved.

A person has a limited time budget or available time to allocate among flexible activities. The sparse spatial distribution and limited durations of activities imply

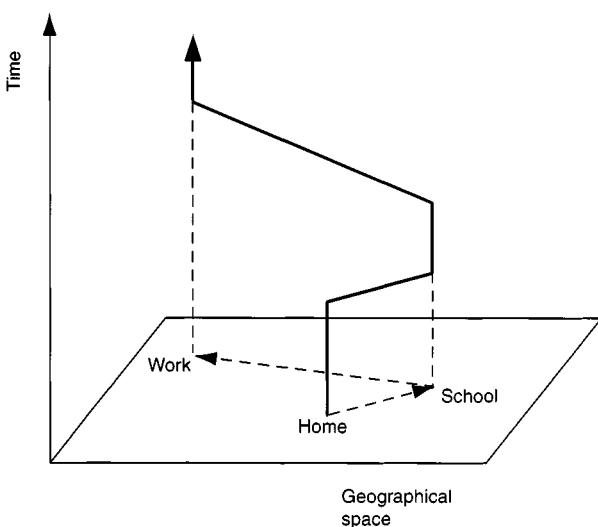


Figure 1. A space–time path.

that the individual must be at different locations at different time periods to participate. This requires the individual to allocate some of their time budget to movement or communication. At a fundamental level, this involves the trading of time for space by the individual. Transportation and communication technologies improve the efficiency of this trade-off by allowing more space to be overcome per unit time.

## 2.2. Space–time path and prism

Although classical time geography recognizes telecommunication, it nevertheless focuses on physical rather than virtual interaction (discussed in more detail later in this chapter). Two central time-geographic concepts are the space–time path and prism. The space–time path traces the individual's physical movement in space with respect to time. Figure 1 provides an example (Miller 2003). The path highlights the constraining effects of a person's need to be at different locations at different times. It also highlights the role of transportation in mitigating these constraints. The slope of the curve illustrates the relationship between time and space in movement. A steeper slope indicates less efficiency in trading time for space, i.e. more time required per unit space in movement. The path is vertical when the individual is stationary in space. The space–time path can be applied at any temporal scale from real-time to a lifespan (Hägerstrand, 1970).

The space-time prism delimits the possible locations for the space-time path. Figure 2 illustrates a prism. Fixed activities anchor a space-time prism, since (by definition) these allow only one spatial possibility during their duration. For example, the two anchoring locations in Figure 2 could be the person's home (which he or she can leave no earlier than time  $t_i$ ) and work (where he or she must be no later than time  $t_j$ ). At some time during the time interval  $t_{ij} = (t_j - t_i)$  the person must stop at some location to conduct an activity that will require at least  $a$  time units. Finally, the person can move with an average maximum velocity  $v$ . The interior of the prism is the potential path space: this shows the points in space and time that the person could occupy during this travel episode. A person cannot participate in an activity unless its space-time path (reflecting its location and available times) intersects the potential path space to a sufficient degree. The projection of the potential path space to geo-space provides the potential path area: all spatial locations that the person could occupy. A person cannot participate in an activity unless its location falls within the potential path area (ignoring the temporal duration of activities).

The path and prism are the fundamental time geographic measure of individual accessibility, where this is defined as the freedom of activity participation in space and time. The number of paths allowed by a given fixed and flexible activity schedule is a surrogate for accessibility (Lenntorp, 1976). We can also form accessibility measures that are a function of the prism size, or the activities and participation times allowed by a prism (Burns, 1979; Miller, 1999).

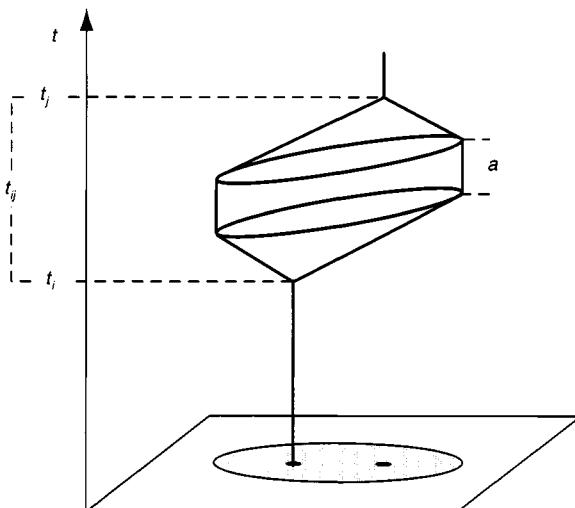


Figure 2. A space-time prism.

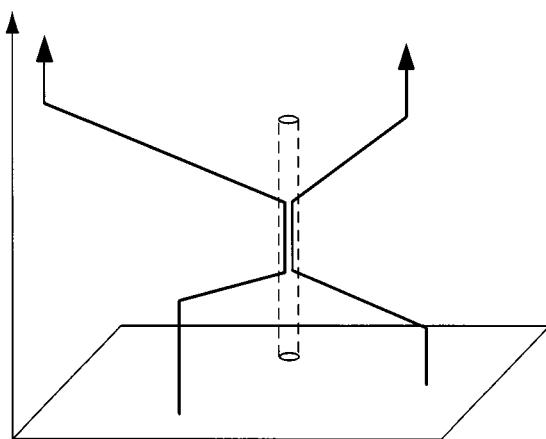


Figure 3. Space-time bundling.

### *2.3. The individual and space-time aggregates*

Time geography encompasses both the individual in space-time and the emergence of space-time aggregates created by the mutual interactions and adjustments of space-time trajectories and activities among multiple persons. A space-time project is a feasible ordering of atomic (indivisible) activities in space and time required to complete some overall goal. This can be applied at any scale, including individual, neighborhood, urban, regional, national, or global, as well as to varying social organization such as the family, work, or community. The requirements for project activities to be sequenced and coordinated lead to the formation of space-time bundles or the convergence of space-time paths. Bundling reflects coupling constraints or the need to be coincident in space and time with other individuals to conduct an activity. Bundling often occurs at space-time stations or locations with resources or infrastructure to support the activity. Figure 3 illustrates two space-time paths bundling at a station (represented as a tube). Activity systems emerge from the “ballet of adjustments” in space-time activities among individuals cooperating or competing within a finite environment (Pred, 1977).

### *2.4. Contrasts with time use and activity analysis*

Besides its explicit treatment of space, time geography may appear similar to time use and activity analysis in economics and transportation science. However, there are several key differences (Pred, 1977). Time geography encompasses time-

scales as long as a person's lifetime as well as corresponding movements at the regional, national, and international scales. While these scales may not be immediately useful to transportation researchers, time geography supports conceptual linkages between transportation and longer-term events such as migration and urban growth. Time geography also does not view time as an infinitely malleable resource that can be divided, reconstituted, and allocated at will by individuals: fixed events and travel velocities condition the ability to allocate time among competing activities. Classical time geography is also concerned with the space-time groupings and aggregates created by the interplay of multiple individuals and how these feedback to constrain individual activities in space and time. Finally, time geography explicitly rejects attempts to predict choices: understanding what a person cannot do is more insightful than understanding what the individual will do.

## 2.5. *Time geography and transportation research*

Despite difficulties in applying time geography to real world problems (see below), basic time-geographic principles have proven to be useful in analyzing individual travel choice and aggregate transportation systems. Lenntorp (1976) shows that space-time accessibility measured by the number of possible paths allowed for a particular activity schedule in a given environment can provide insights into basic planning and policy questions. Burns (1979) uses the space-time prism to compare generic strategies for improving an individual's accessibility to activities, including improving transportation efficiency, different network configurations, and scheduling policies such as facility operating hours. Thill and Horowitz (1997a,b) show that incorporating time constraints into spatial choice models can improve their predictive power. Kwan (1998) shows that space-time-constrained accessibility measures provide qualitatively different portrayals of individual accessibility compared with traditional methods such as the Hansen (1959) measure. Janelle et al. (1998) use time geographic principles to determine the space-time ecology of an urban area from activity diary data.

## 3. **Information technologies and the new time geography**

Hägerstrand and his Lund colleagues developed time geography in the 1950s and 1960s, an era characterized by scarce, expensive data and weak computational platforms (although Sweden was unusually gifted with respect to biographic data). Classical time geography is hampered by unrealistic assumptions such as a maximum velocity that is uniform across space and time as well as difficulties in

collecting data on space–time paths and activities. Classical time geography, while recognizing the possibility of telecommunication, focuses more on physical presence than telepresence (the ability to extend one's reach in space and time through communication technologies).

The world has changed since the formative years of time geography. Data has become cheaper and more readily available. Efficient software exists for handling, analyzing, and visualizing massive data sets, including geo-referenced data. The power of computing platforms has increased exponentially, and will continue to do so for at least the next two decades. Increasing deployment and adoption of communications technologies such as the Internet, the mobile phone, and wireless Internet clients, as well as the development of location-based services (content provision based on geographic location in real time) mean that it is increasingly difficult to separate transportation from communication when studying individual and aggregate travel demand.

Researchers are re-examining time geography in light of new developments in information and communication technologies. The result is a new time geography that is more powerful and relevant with respect to application in the real world. This section of the chapter (based on Miller, 2003) reviews recent contributions and research frontiers in the new time geography.

### *3.1. Representation of space–time environments*

GIS manage, analyze, and communicate information about objects and events in geographic space, and permit the representation and analysis of geographic objects and events at detailed levels of spatial (and increasingly, temporal) resolutions. Detailed digital representations of transportation networks allow researchers to relax the restrictive assumptions that the maximum travel velocity is stationary in space and time.

A network time prism (NTP) is a space–time prism defined within a network with static (constant over time) travel times. A simple procedure exists for calculating the potential path tree (PPT): an NTP product that shows accessible nodes in a network based on the prism constraints (Miller, 1991). The PPT resolves only to the nodes in the network: this can leave unrealistic gaps in network coverage, a weakness made more apparent by the development of address-based geo-referencing methods in GIS. The potential network area (PNA) is an extension of the PPT that resolves to arbitrary locations in the network (Miller, 1999). Figure 4 illustrates a PPT and a PNA defined within the north-east Salt Lake City road network in the USA. Figure 4a shows the reachable nodes in the network based on 15 min travel time from the designated location (indicated by a disk) while Figure 4b shows the reachable network locations, given the same origin and time budget. It is also possible to accommodate

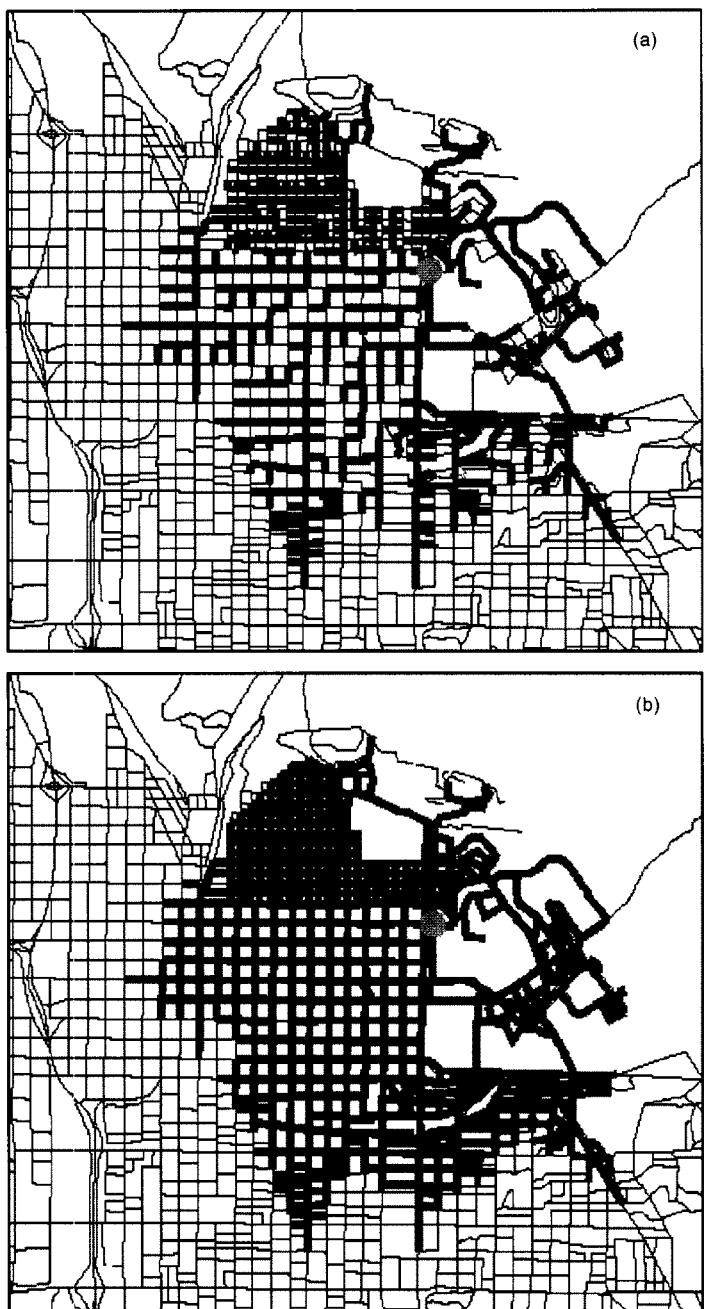


Figure 4. (a) A potential path tree and (b) a potential network area.

heterogeneous networks such as public transit systems with multimodal access (O'Sullivan et al., 2000).

NTPs with static travel times are an improvement over classical time geography but are still unrealistic. Flow and congestion vary over time, creating temporal variations in travel velocities. Congestion also propagates over space and time from localized incidences to wider areas. We cannot use time geography to answer questions about activity timing, “flexitime” scheduling policies, or demand management techniques such as intelligent transportation systems without incorporating time-varying flow and travel velocities. A dynamic network time prism (DNTP) is a space–time prism defined within a network with time-varying velocities and travel times. Wu and Miller (2002) have developed a dynamic potential path tree (DPPT) that shows variations in accessibility in space based on the individual's departure time in a network with discrete-time flow dynamics. Calculating DNTPs in networks with continuous-time dynamics is an open research question.

### *3.2. New methods for data collection*

Space–time activity (STA) data are space–time paths attributed to the activities conducted by the person. Collecting STA data traditionally involved cumbersome methods such as recall and activity diaries. Recall methods require individuals to remember and report activities at a later point in time (say, at the end of the day); this can create errors related to faulty or selective memories. Activity diaries are onerous to participants and suffer from under-reporting errors (Brog et al., 1982; Purvis, 1990). Location-aware technologies (LAT) such as GPS receivers or radiolocation methods that piggyback on wireless communication networks can allow collecting detailed space–time trajectories at the individual level (Smyth, 2001). These space–time trajectories can be presented to individuals to improve recall of activities at a later time (Stopher and Wilmot, 2000).

LATs can also be combined with devices such as personal digital assistants (PDAs) for recording activities directly in digital form. The user-friendliness of PDAs will improve with continuing advances in natural language processing and voice recognition, potentially improving *in situ* activity recording. Current GPS receivers are bulky and consume power at a high rate, and are often linked to vehicles rather than individuals; however, continuing improvements in these technologies will shrink the technology and increase its efficiency – this will allow personal rather than vehicular tracking in space and time to become commonplace. While these advances are a boon to our understanding of transportation and to land use systems, they raise privacy and other ethical concerns. Location privacy methods such as locational masking can help resolve these issues (Armstrong et al., 1999). The effectiveness of locational privacy methods and

the spatio-temporal error induced in time geographic measurements are open research questions.

Location-based services (LBS) refer to information content provision that is sensitive to a user's geographic context, usually through wireless communication devices. LBS are another potential source of STA data since they require tracking of individual space-time paths (Smyth 2001). They also allow the marriage of STA data with an individual's information queries, providing insight on the relationships between individual movement and information access. Using these data will require partnerships with LBS providers; time geographers can offer better support for space-time queries as well as adapting mobile computing services for likely demand patterns across space and time (Miller, 2004).

### *3.3. New methods for data analysis*

A difficulty with STA data analysis is a combinatorial explosion of the information space. Interlinked activity dimensions include the number of activities, and their sequencing, timing, and locations, as well as the transportation/communication modes and possibly routes to access these activities. This implies an information space that is exponential with respect to choice dimensions (Ben-Akiva and Bowman, 1998). Traditional methods for activity analysis require reduction of the information space for tractability. Econometric and statistical approaches and utility-maximizing approaches require *a priori* specification and testing of multidimensional utility functions. Rule-based reasoning systems construct activity and travel schedules based on decision heuristics derived from cognitive science (Garling et al., 1994; Vause, 1997). Simulation methods derive plausible choice sets and simulate individual choices from those sets (Ben-Akiva and Bowman, 1998). All of these techniques can only explore a very small subset of the complex and vast information space of space-time activities.

Advances in information technology for data storage, integration, and analysis can break the combinatorial barrier that has prevented full exploration and discovery of the spatio-temporal patterns in activity data. Data-warehousing techniques are available for integrated and efficient storage of digital geographic data (Bédard et al., 2001); these database design and storage/access techniques must be extended to handle the temporal dimension of STA data. Exploratory visualization techniques for digital geographic data are emerging, but so far only a few techniques are available that can address STA data (Kwan, 2000). Data-mining techniques for STA data include decision tree induction (Arentze et al., 2000) and multidimensional sequencing (Joh et al., 2001).

Still required are effective data warehouse designs to support activity data in space and time. Data warehouses present unique challenges relative to designing transactional databases (operational databases used in daily work processes). For

example, since warehouses need to support fast data retrieval along multiple data dimensions, they require high interconnectivity and redundant storage; these are bad properties for transactional databases with multiple users editing the database on a continual basis (Bédard et al., 2001). Creating effective STA data warehouses requires solving complex issues in spatio-temporal data interoperability, including resolving different semantics, spatial and temporal referencing systems, geometry, accuracy, and precision. Also required is research on standards for activity classification (including aggregation hierarchies) and effective indexing methods for the highly dimensioned STA information space.

### *3.4. Extending time geography to cyberspace*

Although classical time geography recognizes the possibility of telecommunication, it still focuses on physical movement and transportation. Focusing only on physical movement is no longer viable in a world where transportation and telecommunication have altered dramatically the nature of space and time (Janelle, 1969; Couclelis and Getis, 2000). The popular “death of distance” argument is simplistic: information technologies such as the Internet and mobile phones are complements as well as substitutes for transportation, increasing the level and complexity of travel demand. For example, a city guide website may induce people to travel to a neighborhood or city that they may have never visited before. Another example is cell phones creating “flocking” behavior, where social interactions evolve fluidly over the course of an evening as calls are exchanged rather than being pre-planned. Time geography provides a natural fit with emerging perspectives that view time as the scarce commodity within lifestyles and societies accelerated by information technologies (e.g. Gleick, 1999).

Extensibility diagrams are modifications of the space–time path that encompass movement and communication at all geographic scales (Adams, 2000). Extensibility diagrams can illustrate general characteristics of the relationships between transportation and communication in activity participation. Individuals can be compared with respect to the frequency, duration, time, and geographic scale of travel, incoming communication, and outgoing communication. However, geographic space is restricted to only a crude ordinal scale (local, regional, national, etc.). Still needed are tools with higher spatial resolution and the ability to support synoptic summaries among multiple individuals, as well as drill-down analysis and other exploratory and data-mining techniques.

Classical time geography also ignores the use of communication and information technologies to reduce uncertainty through information search and learning. Uncertainty about transportation system performance (i.e. travel velocity) can induce earlier departures, while uncertainty about the locations

and attributes of activities can induce search behavior. Both can reduce accessibility by reducing the time available for other activities (Hall, 1983). Kwan and Hong (1998) integrate cognitive constraints (e.g. preferences or lack of information) into an NTP through an effective but *ad hoc* overlay procedure. An extended research effort is required that re-examines time geography from its foundation and reformulates an analytical theory (similar to Burns, 1979) that recognizes imperfect information.

#### 4. Conclusion

Transportation and communication technologies exist to alter the relationship between geographic space and time, making it feasible for individuals to participate in activities in more locations and times. Because space and time are central to human existence, these technologies profoundly impact our lives and the world (Janelle, 1969). Time geography is an elegant perspective for analyzing the interrelationships among activities in space and time, and the role of transportation and communication technologies in facilitating and constraining these interrelationships.

While basic time-geographic principles have proven to be useful in travel behavior and transportation research, the potential power of time geography is being enhanced tremendously by the rise of geographic information technologies such as GIS and LAT. Geo-information technologies allow more detailed representation of space-time environments, easier and more accurate space-time activity data collection, and computationally scalable methods for discovering new knowledge from massive space-time activity data sets. The rise of new information technologies is also changing the real-world basis for time geography by blurring the boundaries between transportation and communication. A new time geography is emerging that has tremendous potential as well as imperative and fascinating research challenges.

#### References

- Adams, P.C. (2000) "Application of a CAD-based accessibility model," in: D.G. Janelle and D.C. Hodge, eds, *Information, place and cyberspace: issues in accessibility*. Berlin: Springer-Verlag.
- Arentze, T.A., F. Hofman, H. van Mourik, H.J.P. Timmermans and G. Wets (2000) "Using decision tree induction systems for modeling space-time behavior," *Geographical Analysis*, 32:330-350.
- Armstrong, M.P., G. Rushton and D.L. Zimmerman (1999) "Geographically masking health data to preserve confidentiality," *Statistics in Medicine*, 18:497-525.
- Bédard, Y., T Merrett and J. Han (2001) "Fundamentals of spatial data warehousing for geographic knowledge discovery," in: H.J. Miller and J. Han, eds, *Geographic data mining and knowledge discovery*. London: Taylor and Francis.

- Ben-Akiva, M.E. and M.E. Bowman (1998) "Activity based travel demand model systems," in: P. Marcotte and S. Nguyen, eds, *Equilibrium and advanced transportation modeling*. Boston: Kluwer.
- Brog, W., E. Erl, A.H. Meyburg and M.J. Wermuth (1982) "Problems of non-reported trips in surveys of nonhome activity patterns," *Transportations Research Record*, 891:1-5.
- Burns, L.D. (1979) *Transportation, temporal and spatial components of accessibility*. Lexington: Lexington Books.
- Couclelis, H. and A. Getis (2000) "Conceptualizing and measuring accessibility within physical and virtual spaces," in D.G. Janelle and D.C. Hodge, eds, *Information, place and cyberspace: issues in accessibility*. Berlin: Springer-Verlag.
- Garling, T., M.-P. Kwan and R.G. Golledge (1994) "Computational-process modeling of household activity scheduling," *Transportation Research B*, 26:355-364.
- Gleick, J. (1999) *Faster: the acceleration of just about everything*. New York: Pantheon Books.
- Hägerstrand, T. (1970) "What about people in regional science?" *Papers of the Regional Science Association*, 24:7-21.
- Hall, R.W. (1983) "Travel outcome and performance: the effect of uncertainty on accessibility," *Transportation Research B*, 17:275-290.
- Hansen, W.G. (1959) "How accessibility shapes land use," *Journal of the American Institute of Planners*, 25:73-76.
- Janelle, D.G. (1969) "Spatial organization: a model and concept," *Annals of the Association of American Geographers*, 59:348-364.
- Janelle, D.G., M.F. Goodchild and B. Klinkenberg (1998) "The temporal ordering of urban space and daily activity patterns for population role groups," *Geographical Systems*, 5:117-137.
- Joh, C.-H., T.A. Arentze and H.J.P. Timmermans (2001) "Multidimensional sequence alignment methods for activity-travel pattern analysis: a comparison of dynamic programming and genetic algorithms," *Geographical Analysis*, 33:247-270.
- Kwan, M.-P. (1998) "Space-time and integral measures of accessibility: a comparative analysis using a point-based framework," *Geographical Analysis*, 30:191-216.
- Kwan, M.-P. (2000) "Interactive geovisualization of activity-travel patterns using three-dimensional geographical information systems: a methodological exploration with a large data set," *Transportation Research C*, 8:185-203.
- Kwan, M.-P. and X.-D. Hong (1998) "Network-based constraints-oriented choice set formation using GIS," *Geographical Systems*, 5:139-162.
- Lenntorp, B. (1976) *Paths in space-time environments: a time geographic study of movement possibilities of individuals*. Lund studies in geography number 44. Lund: Royal University of Lund.
- Miller, H.J. (1991) "Modeling accessibility using space-time prism concepts within geographical information systems," *International Journal of Geographical Information Systems*, 5:287-301.
- Miller, H.J. (1999) "Measuring space-time accessibility benefits within transportation networks: basic theory and computational methods," *Geographical Analysis*, 31:187-212.
- Miller, H.J. (2004) "What about people in geographic information science?" in: D. Unwin, ed., *Representing geographic information systems*. Chichester: Wiley.
- O'Sullivan, D., A. Morrison and J. Shearer (2000) "Using desktop GIS for the investigation of accessibility by public transport: an isochrone approach," *International Journal of Geographical Information Science*, 14:85-104.
- Pred, A. (1977) "The choreography of existence: comments on Hagerstrand's time-geography and its usefulness," *Economic Geography*, 53:207-221.
- Purvis, C.L. (1990) "Survey of travel surveys II," *Transportation Research Record*, 1271:23-32.
- Smyth, C.S. (2001) "Mining mobile trajectories," in: H.J. Miller and J. Han, eds, *Geographic data mining and knowledge discovery*. London: Taylor and Francis.
- Stopher, P.R. and C.G. Wilmot (2000) "Some new approaches to designing household travel surveys: time-use diaries and GPS," in: *Proceedings of the 79th Annual Meeting of the Transportation Research Board*. Washington, DC.
- Thill, J.-C. and J.L. Horowitz (1997a) "Modelling non-work destination choices with choice sets defined by travel-time constraints," in: M.M. Fischer and A. Getis, eds, *Recent developments in spatial analysis: spatial statistics, behavioural modelling and computational intelligence*. Berlin: Springer-Verlag.

- Thill, J.-C. and J.L. Horowitz (1997b) "Travel time constraints on destination-choice sets," *Geographical Analysis*, 29:108–123.
- Vause, M. (1997) "A rule-based model of activity scheduling behavior," in: D.F. Ettema and H.J.P. Timmermans, eds, *Activity-based approaches to travel analysis*. Oxford: Elsevier.
- Wu, Y.-H. and Miller, H.J. (2002) "Computational tools for measuring space–time accessibility within transportation networks with dynamic flow," *Journal of Transportation and Statistics*, 4:1–14.

# AUTHOR INDEX

---

## Index Terms

## Links

### **A**

Abdulaal, M.S.	567	570		
Abraham, J.E.	132	201		
Acs, Z.J.	36			
Adams, P.C.	657			
Adams, T.M.	325			
Adler, T.	386	521	527	
Allen, G.L.	533	534		
Allgower, E.L.	276			
Allsop, R.E.	572			
Alonso, W.	1	28	134	168
	203			
Amin, A.	33	132	134	147
	201	241		
Amiri, E.	496			
Anderson, J.	551			
Anderson, M.D.	344	352		
Anderstig, C.	131			
Anderton, D.L.	48			
Ando, A.	278			
Apgar, W.C. Jr	140			
Appleyard, D.	544			
Arentze, T.	224	510	526	527
	656			
Armington, P.S.	273			

This page has been reformatted by Knovel to provide easier navigation.

## Index Terms

Armstrong, M.P.	47	655
Aschauer, D.A.	77	94
Assad, A.A.	400	403
Atzeni, P.	393	
Axhausen, K.W.	140	

## **B**

Baber, C.M.	384			
Bachu, P.	388	437		
Badoe, D.A.	149			
Bae, C.-H.C.	259	261	263	
Bagley, M.N.	262			
Bak, P.	250			
Bandy, G.	384			
Bania, N.	74			
Banister, D.	262	263	583	
Baraniak, D.	493			
Bar-Gera, H.	584			
Barrett, C.L.	554	555	556	557
	562			
Barro, R.J.	284			
Bass, T.	584			
Batten, D.	241	242	249	250
Batty, M.	140	185	551	557
	561			
Batty, P.	428			
Beckman, R.J.	140	220		
Beckmann, M.J.	516	585	590	
Bédard, Y.	656	657		
Been, V.	49			

## Links

This page has been reformatted by Knovel to provide easier navigation.

## Index Terms

## Links

Bell, M.G.H.	392	567	568	611
	615	620		
Ben-Akiva, M.E.	140	149	224	230
	241	516	521	527
	615	618	656	
Benati, S.		224		
Benekohala, R.F.		618		
Benenson, I.		551		
Benz, R.J.		470		
Berechman, J.		128		
Bertaud, A.		255		
Bertsekas, D.P.		585		
Bespalko, S.J.		399		
Bhat, C.R.	19	618	623	
Biehl, D.	77	94		
Birch, E.		260		
Black, A.		115		
Black, W.R.		20		
Blumberg, E.		74		
Boarnet, M.	258	259	260	262
		263		
Bodin, L.	358	403		
Bollens, S.A.		30		
Bollinger, C.		260		
Bovy, P.H.L.		559		
Bowen, W.M.	43	44	45	46
	47	49	55	
Bowlby, S.R.		520		
Bowman, J.		527		
Bowman, M.E.		656		

## Index Terms

## Links

Boyce, D.E.	128	131	241	585
	600	604	609	611
Braden, B.M.	59			
Braess, D.	584			
Britch, S.C.	491	492	497	
Brock, W.A.	247			
Brocker, J.	188	277	278	280
	283	285		
Brog, W.	655			
Brotchie, J.F.	135			
Brown, D.L.	30			
Brown, P.	50			
Brueckner, J.J.	81			
Bryant, B.	44	46	51	
Buckley, P.H.	278	279		
Bullard, R.D.	46	47		
Bullock, D.	471	474	476	477
	478			
Bullock, P.	386	388	515	519
Burns, L.D.	650	658		
Burrough, P.A.	310			
Burtch, R.	493			
Bush, B.W.	562			
Butler, S.E.	87			
Button, K.J.	86	89	92	367
	583			

## C

Calthorpe, P.	261
Cao, B.	359

This page has been reformatted by Knovel to provide easier navigation.

## Index Terms

## Links

Capozza, D.	260				
Cascetta, E.	614	615	620	623	
Cassir, C.	611				
Casti, J.	244				
Cervero, R.	72	75	260		
Cetin, N.	555	557			
Chakraborty, J.	47				
Chan, K.S.	610	617	620	623	
Chapin, F.S.	139	514	630	632	
Chatman, D.	74	263			
Chen, D.	50				
Chen, K.	250				
Chen, P.S.T.	508				
Choi, Y.L.	609				
Chorley, R.	14				
Christaller, W.	2	28			
Clark, S.D.	623				
Clarke, G.P.	140				
Clarke, K.	224				
Clarke, M.	140				
Clingermayer, J.C.	33				
Cohen, J.	584				
Collins, R.W.	46				
Compin, N.	260				
Conrad, K.	278				
Corburn, J.	59				
Couclelis, H., R G.	224	503	657		
Coursey, D.L.	59				
Coyle, M.	48				

## Index Terms

## Links

Crane, R.	74	258	262	263
	264			
Cullen, I.G.	630			
Cutter, S.L.	44	47		
Czamanski, S.	98			
Czerniak, R.	427	494		

## **D**

Dafermos, S.C.	585	586	589	591
	594	595	598	600
	601	603	611	615
Daganzo, C.F.	595			
Dane, C.	399			
Dantzig, G.B.	402			
De Gruyter, C.	388			
de la Barra, T.	132	147	200	204
	224	615		
Deakin, E.	336			
Deming, W.	227			
Deng, F.F.	40			
Dennis, N.P.S.	79			
Desrosiers, J.	405	406		
Deutschman, H.D.	297			
Dial, R.B.	595	600	614	615
Dickey, J.W.	223			
Diewert, W.E.	275			
Dijst, M.	519			
DiPasquale, D.	231			
Divis, D.A.	417	426		
Dixit, A.K.	283	285		

This page has been reformatted by Knovel to provide easier navigation.

## Index Terms

## Links

Doddridge, J.	541	542		
Doganis, R.	79			
Doherty, S.T.	385	388		
Domencich, T.A.	134	186	240	241
Dong, J.	584	587	592	599
	600	604	606	
Donoso, P.P.	132	147		
Doupe, P.J.	465			
Downs, R.M.	536	537		
Draijer, G.	435			
Drane, C.	494			
Dresner, M.	91			
Drummond, W.J.	380			
Duany, A.	261			
Dueker, K.J.	329			
Dupuis, P.	595	604	605	

## **E**

Easterling, D.	45			
Echenique, M.H.	132	187	188	200
Ellegård, K.	525			
Elliott, D.H.	632			
Ellis, C.D.	492			
Elmi, A.M.	516			
Engelen, G.	224			
Eppli, M.J.	261			
Ettema, D.F.	514	526	637	
Ewing, R.	30			

## Index Terms

## Links

### F

Faghri, A.	471				
Feiock, R.C.	33				
Ferrand, N.	140				
Fischer, M.M.	247	400	403	404	
Fisk, C.	614				
Fitch, G.M.	491	492	497		
Flood, M.	385				
Florian, M.	585	594	598		
Flowerdew, R.	335				
Fohl, P.	395	399			
Forkenbrock, D.J.	50				
Fotheringham, A.S.	517				
Frank, M.	594				
Freeman, M.	48	49			
Freundschatz, S.	534	536			
Friedman, B.	212				
Friesz, T L.	281	598			
Fujii, S.	526				
Fujita, M.P.	222	283			
Fulton, W.	255				
Fürst, F.	128				

### G

Gale, N.D.	508				
Gallager, R.	586				
Gangrade, S.	524				
Garin, R.A.	177				

## Index Terms

## Links

Gärling, T.A.	140	501	506	508
	509	510	533	535
	536	538	656	
Garreau, J.	28	31	257	
Garrett, M.	74	204		
Garrido, R.A.	618			
Garrison, W.L.	17	22		
Gasiorek, M.	278			
Gaydos, L.	224			
Gazel, R.	278			
Genosko, J.	278	282		
Georg, K.	276			
Geraldes, P.	200			
Getis, A.	657			
Gilks, W.R.	231			
Gillingwater, D.	407			
Gimblett, H.R.	550	551		
Giuliano, G.	257	265		
Gleick, J.	657			
Glover, E.	402			
Goetz, A.R.	18			
Golden, G.B.	400	403		
Goldner, W.	203			
Golledge, R.G.	501	502	506	509
	535	536	538	
Golob, J.M.	519			
Golob, T.E.	516	519	521	522
Goodchild, M.E.	391	393	394	395
	396	398	399	631
	632			

## Index Terms

## Links

Gordon	29	73		
Goulias, K.G.	521	524		
Graettinger, A.J.	496			
Grantham, C.E.	30			
Gray, R.G.	24			
Greaves, S.P.	376	382	383	388
	389	430		
Green, M.B.	98	102	335	
Greenberg, M.	43			
Greene, W.	230	231		
Guensler, R.	435			
Gunn, H.E.	526			
Guo, P.	471			
Gurney, K.	550			

## **H**

Haag, G.	132	134	246	
Haddad, E.	278			
Hägerstrand, T.	139	293	630	647
	648	649		
Haggett, P.	14			
Haklay, M.	559	560	561	
Hall, M.D.	609			
Hall, R.W.	658			
Halseth, G.	541	542		
Hamad, K.	471			
Hamerslag, R.	585			
Hamilton, J.T.	47	48		
Hansen, W.G.	129	652		
Hanson, P.	523			

This page has been reformatted by Knovel to provide easier navigation.

## Index Terms

## Links

Hanson, S.	515	517	519	523
Harberger, A.C.	269			
Harker, P.T.	598			
Harner, J.	59			
Harper, E.	321			
Harris, B.	128			
Harris, C.D.	28			
Harvey, A.S.	637			
Harvey, G.	336			
Haynes, K.E.	44	46	51	55
Helbing, D.	560	561		
Hemmens, G.C.	514	525		
Heng, St.	278			
Hensher, D.	132	147	247	
Hepworth, T.	396			
Hertel, T.W.	269	284		
Hewings, G.J.D.	278			
Hickman, R.	262	263		
Hightower, H.C.	632			
Hird, J.	44	48		
Hirsh, M.	523			
Hirte, G.	278			
Holland, J.H.	550			
Holling, C.S.	248			
Holzer, H.J.	72			
Hong, X.-D.	658			
Hoogendorn, S.P.	559			
Hook, P.	460	461	464	
Horgan, P.	244			
Horner, M.	352			

## Index Terms

## Links

Horowitz, J.L.	652
Horridge, M.	280
Horton, F.E.	519      520      522      631
Hounsell, N.	456
Hoyt, H.	28
Huang, H.J.	611
Huff, J.O.	523
Hughes, P.C.	620
Hultquist, J.F.	522
Hunt, J.D.	132      147      201
Hussain, I.	279      282

## **I**

Ihlantfeldt, K.R.	72      260
Iida, Y.	392
Ingram, G.K.	265
Ishikawa, T.	506      508

## **J**

Jacobson, J.	524
Jaikumar, R.	403      404
Janelle, D.G.	631      652      657      658
Javeau, C.	630      631      632
Jensen, H.J.	249      250
Joh, C.-H.	521      523      524      656
Johansen, L.	269
Johnson, D.S.	402
Johnston, R.A.	224

## Index Terms

Jones, P.M.	385	514	633	637
	644			
Jones, R.	278			
Jou, R.-C.	521			
Juan, H.E.	461			
Jun, M.-J.	263			

## **K**

Kahn, P.	212			
Kain, J.R.	72	140		
Kansky, K.J.	522			
Kaplan, E.D.	413	416	423	
Karimi, H.A.	493			
Karr, D.	496			
Kassof, H.	297			
Kauffman, S.A.	250			
Kellerman, A.	30			
Kelly, F.P.	584			
Kennedy, L.G.	49			
Kenworthy, J.	264	265		
Kernighan, B.	402			
Kerridge, J.	559	560	561	
Keuleers, B.	524			
Khattak, A.J.	72			
Khoo, V.H.S.	451			
Kiernan, L.J.	87			
Kim, T.-G.	524			
Kim, T.J.	131			
Kinderlehrer, D.	594			
Kirkpatrick, S.	402			

## Links

This page has been reformatted by Knovel to provide easier navigation.

## Index Terms

## Links

Kitamura, R.	518	520	521	524
	525	526		
Kitchin, R.M.	501	534	536	
Klosterman, R.E.	224			
Knaap, T.	280			
Knight, F.H.	583			
Knight, R.L.	150			
Koenig, J.G.	515			
Kofoed, J.	516			
Kolata, G.	584			
Koller, W.	247			
Kondo, K.	519			
Koniditsiotis, C.	455			
Koppelman, F.S.	14	526		
Korilis, Y.A.	584			
Koyama, N.	633			
Kreibich, V.	140			
Kreitz, M.	375	385		
Krugman, P.	278	283		
Kuby, M.J.	24			
Kulmala, R.	456			
Kunreuther, H.	45			
Kurose, S.	561			
Kurzweil, R.	550			
Kut, S.	493			
Kwan, M.-P.	23	395	510	652
	656	658		

## Index Terms

## Links

### **L**

Laird, D.	469			
Lakshmanan, T.R.	396			
Lalanne, L.	14			
Lall, S.	89			
Lam, W.H.K.	567	571	610	615
	617	620	623	
Landis, J.D.	131	140	224	335
Laporte, G.	401			
Lavelle, M.	48			
Lawler, E.L.	400			
LeBlance, L.J.	567	570		
Lee, D.B.	185	205	337	
Lehtonen, M.	456			
Leinbach, T.R.	22			
Leiner, C.	223			
Lemos, R.	455	465		
Lenntorp, B.	293	514	525	650
	651			
Leontief, W.W.	187	189	203	223
Lerman, S.R.	224	230	241	516
	520			
Lette, L.	74			
Leurent, F.	600			
Levy, S.	550			
Ley, D.	542			
Li, T.Y.	242			
Liben, L.S.	505			
Lin, S.	402			

## Index Terms

## Links

Lin Salem, P.	92
Lingsom, S.	637
Lipset, S.M.	35      41
Liu, Z.	265
Llewellyn, L.	49
Lloyd, R.	533
Lo, H.	611      616
Lomax, T.J.	69      452      470      476
Longfoot, J.	452
Longley, P.A.	391      393      397      398 402
Lös, M.	598
Losch, A.	2
Lowry, I.S.	131      167      185      580
Luce, R.D.	615
Luk, J.Y.K.	451      461
Luoma, M.	516
Lynch, K.	537      543

## **M**

Maat, C.	367      368
Mackett, R.L.	128      131      223
Madden, J.R.	278
Maher, M.J.	620
Mahmassani, H.S.	457      508      521      618
Malpezzi, S.	255
Mammano, F.J.	458
Mamuneas	21
Mannermaa, M.	244
Marble, D.F.	519      520

This page has been reformatted by Knovel to provide easier navigation.

## Index Terms

## Links

Marcotte, P.	585
Martellato, D.	244
Martin, J.	460
Martínez, F.J.	132      147
Masters, E.G.	492
Matei, S.	541      542
Mattsson, L.-G.	131
May, R.	241
Mayer, C.	79
Mayeres, I.	278
McCarthy, G.M.	297
McDonald, K.	416
McFadden, D.	134      186      187      224 240      241      509
McLafferty, S.	72
McLellan, J.F.	495
McNally, M.G.	523
McNaughton, R.B.	98      102
Medda, F.	259      263
Meng, Q.	573      580
Metzger, J.	49
Meyburg, A.H.	294
Meyer, M.D.	112
Migdalas, A.	567
Mikkonen, K.	516
Miller, E.J.	112      131      140      141 147      149
Miller, H.J.	23      393      394      397 398      399      550      649. 653      655      656

This page has been reformatted by Knovel to provide easier navigation.

## Index Terms

## Links

Miller, J.S.	496
Milligan, J.	29
Mills, E.S.	203
Mills, G.S.	49
Mitchell, J.T.	59
Miyamoto, K.	132
Mokhtarian, P.L.	24      262
Moore, T.	213
Morris, J.	412
Muller, P.O.	256
Murakami, E.	72      386
Muth, R.F.	203

## **N**

Nadiri, M.I.	21
Nagel, K.	140      554      562
Nagurney, A.	584      587      591      592
	594      595      598      599
	600      603      604      605
	606      613
Narayan, D.	265
Nelson, A.C.	112
Neuhäuser, K.S.	49
Newman, P.	264      265
Ng, J.C.	384
Nguyen, S.	585
Nichols, L.D.	30
Nicolis, G.	242      249
Nicot, B.-H.	255
Nielsen, A.	618

This page has been reformatted by Knovel to provide easier navigation.

## Index Terms

## Links

Nijkamp, P.	20	238	239	241
	242	243	244	245
	247	251		
Niles, J.M.		30		
Nivola, P.S.		259		
Noronha, V.		490		
North, D.	33	34		
Noth, M.		206		
Nuttall, I.		458		
Nyerges, T.L.		397		
Nystuen, J.D.		519		

## O

O'Brien, T.V.	83			
O'Kelly, M.E.	23	520		
O'Neill, W.	321			
Oehry, B.	461	462		
Oei, H.L.	459	460		
Ogden, K.W.	465			
Ogden, M.A.	470			
Okamoto, M.	458			
Oloufa, A.A.	456			
Olurotimi, O.	247			
Ong, P.	74			
Oosterhaven, J.	280			
Oppenheim, N.	522			
Orcutt, G.	139	223		
Ortuzar, J. de D.	239			
Ostrpwski, P.L.	83			
O'Sullivan, D.	561	655		

This page has been reformatted by Knovel to provide easier navigation.

## Index Terms

## Links

### P

Paelinck, J.H.P.	99	
Park, R.E.	28	
Pas, E.I.	523	524
Patriksson, M.	568	570
Penn, A.	560	
Perez, M.	471	479
Perlin, S.A.	48	
Perrings, C.	248	
Peters, J.I.	458	
Pfaffenbichler, P.C.	135	
Phelps, E.	630	
Pigou, A. C.	583	
Pimm, S.L.	248	
Pisarski, A.	73	
Polak, P.H.	459	460
Poling, A.D.	471	492
Pollack, H.	44	48
Polydoropoulou, A.	465	
Potvin, J.-Y.	401	402
Powell, W.B.	618	
Prastacos, P.	132	223
Pred, A.	648	651
Preston, V.	72	
Prigogine, I.	242	249
Prud'homme, R.	255	
Pucher, J.	257	
Pucher, M.	69	
Pulido, L.	50	

## Index Terms

## Links

Purvis, C.L.	655
Pushkarev, B.S.	119
Putman, S.H.	131      172      203      222

## **Q**

Quandt, R.E.	599
Quercia, R.	72
Quiroga, C.A.	471      474      476      477 478      479

## **R**

Rabah, M.	457
Rabin, Y.	49
Ran, B.	584      604      611
Recker, W.W.	523      526
Reggiani, A.	238      239      241      242 243      244      245      248 249      251
Reilly, J.R.	427      494
Reismann, M.	247
Renne, J.L.	69      257
Reynolds, D.R.	631
Rho, J.H.	131
Richardson, H.W.	29      257
Rickert, M.	555      556      557      562
Rizos, C.	399      494
Rockafellar, R.T.	568
Rohr, C.	200
Rusk, D.	30

## Index Terms

## Links

### S

Sala-I-Martin, X.	284			
Salomon, I.	136	140	214	
Salvini, P.A.	131			
Samuelson, P.A.	278	585	598	
Sanchez, T.W.	121			
Sarjeant, P.M.	384			
Schafer, J.	227			
Schintler, L.A.	247			
Schleiffer, R.	500			
Schneider, M.	599			
Schrank, D.L.	69	470		
Schrekenberg, M.	554			
Schweitzer, L.A.	50			
Sen, A.	241			
Shaw, S.-L.	13	393	394	397
	398	399	550	
Sheffi, Y.	196	281	570	595
	609	610	611	613
	615	618		
Shen, Q.	72	118		
Shepherd, S.P.	135			
Shibata, T.	278			
Shiftan, Y.	262			
Sholl, M.J.	504			
Shoven, J.B.	269			
Shunk, G.A.	325			
Siaurusaitis, V.	385			
Simmonds, D.C.	131	132	147	226

## Index Terms

## Links

Sinai, T.	79
Sipper, M.	550
Sjoquist, D.L.	72
Slavin, H.	355
Small, K.	128
Smith, T.R.	510
Smith, T.E.	241
Smyth, C.S.	655
Snellen, D.	656
Somers, A.	524
Souleyrette, R.R.	430
Southworth, F.	22
Sparrow, F.T.	344
Spear, B.D.	147
Spiekermann, K.	586
Spiess, H.	589
Spiller, P.T.	594
Stampacchia, G.	396
Stanback, T.	594
Stea, D.	30
Steiner, M.	536
Stephan, F.	537
Steiner, M.	102
Stephan, F.	227
Stiglitz, J.E.	283
Stimson, R.J.	285
Stopher, P.R.	536
Stough, R.	294
	353
	375
	386
	388
	433
	435
	441
	443
	633
	637
	638
	655
Stough, R.	30
	31
	32
	99
	101

This page has been reformatted by Knovel to provide easier navigation.

## Index Terms

## Links

Stradling, S.	460			
Strathman, J.G.	520	522		
Stretcher, C.	383			
Stutz, F.	44			
Suhrbier, J.	262			
Sundberg, J.	459			
Suppes, P.	615			
Sutton, J.	396	407		
Swarm, G.D.	225			
Szalai, A.	634	636		
Szeto, Y.W.	611			

## T

Taniguchi, E.	364	369	371		
Taylor, B.	74				
Taylor, D.E.	45				
Taylor, M.A.P.	471	585			
Taylor, S.Y.	92				
Tesfatsion, L.	225				
Thill, J.-C.	23	550	521	651	
Thittai, R.	425				
Thomas, I.	521				
Thorsen, S.	439				
Thrift, N.	33				
Timmermans, H.P.J.	510	514	515	521	
	525	526			
Tomlinson, J.	525				
Ton, T.T.	132	247	329		
Torrens, P.M.	549	551	561		
Trela, I.	278				

## Index Terms

## Links

Tretheway, M.W.	91
Trygg, L.L.	150
Tu, C.C.	261
Tung, Y.K.	611
Turner, A.	560
Turner, S.	476
Tversky, B.	536      538

## **U**

Udomsri, R.	132
Ueda, T.	278      280
Ueno, M.	435

## **V**

Valentini, M.P.	367
Van der Hoorn, T.	526
Vause, M.	656
Venables, A.J.	278      282
Vidakovic, V.	519
Visser, J.G.	367      368
Vittas, E.M.	44      48
Volmuller, J.	585
von Rosenbladt, B.	632
von Thunen, J.H.	28      77      168      278
Vonderohe, A.	396

## Index Terms

## Links

### **W**

Wachs, M.	204	515		
Waddell, P.	132	140	147	206
	207	209	213	217
	223	228	232	335
Wagner, D.P.	386	433		
Wagner, P.	555	556	557	
Wagner, W.E.	519	520		
Wales, T.J.	275			
Wall, G.	456			
Wallace, B.	524			
Walz, U.	284			
Wang, M.S.	32			
Ward, J.D.	17			
Wardrop, J.G.	585	610	612	
Wardwell, J.M.	30			
Waters, N.	338			
Watling, D.P.	623			
Webster, F.V.	185			
Wegener, M.	128	131	137	139
	140	147	157	185
	223	335		
Weiner, E.	203	310		
Weiss, S.F.	139			
Wellington, A.M.	14			
Wen, C.-H.	526			
Wermuth, M.	385			
Westin, L.	282			
Wets, G.	524			

## Index Terms

## Links

Whalley, J.	269	278		
Wheaton, W.C.	213			
Wheeler, J.O.	520			
White, R.	224			
Wiegel, D.	359			
Williams, H.C.W.L.	186	187	193	196
Williams, I.N.	199	200		
Williamson, O.E.	28	33	34	35
Willumsen, L.G.	239	609		
Wilmot, C.G.	633	638	655	
Wilson, A.G.	14	128	134	170
	186	239	249	516
Wilson, G.W.	22			
Winston, C.	94			
Wolf, A.	247			
Wolf, J.	388	430	425	534
Wolfe, P.	594			
Wolfram, S.	224			
Wong, S.C.	567	571		
Woods, R.	494			
Wright, B.H.	50	55		
Wu, J.-H.	584	655		
Wylie, B.	555	557		

## X

Xie, Y.	140
Xu, G.	620

## Index Terms

## Links

### **Y**

Yamamoto, T.	527				
Yang, H.	567	567	571	573	
	580	610			
Yang, Q.	355				
Ygnace, J.	452				
Yorke, J.A.	242				
Young, J.	72				

### **Z**

Zhang, D.	595	598	604	605	
	606				
Zhang, M.	131	140			
Zhaob, W.	618				
Zimmerman, R.	48				

# SUBJECT INDEX

---

## Index Terms

## Links

### A

accessibility	121	136	260	515
accidents, <i>see</i> safety				
activity analysis	19	142	240	353
	386	514	554	560
	629	647		
acts				
Airline Deregulation Act (1978)	18			
Civil Rights Act (1964)	46			
Clean Air Act (1990)	138			
Clean Air Amendments Act (1991)	294	211		
Emergency Planning and Community Right-to-know Act (1986)	45			
Federal-aid Highway Act (1956)	14			
Growth Management Act	207	217		
Intermodal Surface Transportation Efficiency Act (1991)	23	138	204	
National Environmental Policy Act (1969)	21	497		
Rail Revitalization and Regulatory Reform Act (1976)	18			
Regional Rail Reorganization Act (1973)	18			
Staggers Rail Act (1980)	18			
Super Fund Amendments and Reauthorization Act (1986)	45			
Transportation Equity Act (1998)	138			

This page has been reformatted by Knovel to provide easier navigation.

<u>Index Terms</u>	<u>Links</u>			
advanced driver assistance systems	460			
advanced traveler information system	508			
agent-based models	526      558      560			
agglomeration economies	29      169      263			
agriculture	128			
Air Transport Action Group	89			
Airborne	24			
airlines	18      78      218      414			
	417      503			
airports	78      316			
American Housing Survey	263			
Association of State Highway and Transportation Officials	313			
automatic vehicle location	456      490      493			
axial cities	28			
<b>B</b>				
barges	218			
benefit-cost analysis, <i>see</i> cost-benefit analysis				
bicycles	213      261      433      440			
	526      639			
bid-rent curves	134      168			
<i>see also</i> rent				
Braess paradox	584			
buses	22      68      74      119			
	196      211      213      217			
	257      263      265      283			
	302      316      320      456			
	503      528			
business travel	280			

This page has been reformatted by Knovel to provide easier navigation.

## Index Terms

## Links

### C

Caliper Corporation	23	338	361
Cambridge Systematics	21		
car ownership/use	121	148	150
	257	258	263
	433	469	503
	528	551	556
	639		
car pooling	204	213	
cellular automata	549	552	558
cellular dynamics	140		
Census Transportation Planning Package	312		
central place theory	2	28	
Channel Tunnel	200		
Chicago Convention	78		
Cobb–Douglas utility function	184		
cognitive maps	501	533	
Commission for Racial Justice	45	48	
common interest developments	40		
commuting	71	111	116
	255	280	149
distance	265		
competition			
imperfect	279	283	
perfect	270	283	
complexity theory	244		
concentric cities	2	28	

## Index Terms

## Links

congestion	3	17	38	68
	79	81	112	197
	199	201	265	281
	302	369	430	461
	507	511	554	556
	567	583	610	616
connectivity	121	394		
consumers' surplus	272	276		
containerization		17		
cost minimization	27	507		
cost–benefit analysis	199	210	270	
Council of Logistics Management		357		
cycling, <i>see</i> bicycles				

## **D**

data issues	294	309	330	
<i>see also</i> geographical information systems				
demand management	161			
demographics	155			
deregulation	18			
dial-a-ride	457			
disaggregate models	172	187	196	200
	239			
discrete choice models	224	230	237	340
<i>see also</i> logit models and probit models				
distance measuring instruments	470	487	493	

## Index Terms

## Links

### **E**

economies				
density	81			
scale	77	79	101	
scope	77	79	101	
edge cities	29	31	257	
elasticities, price	81	282	596	598
employment	71	75	79	90
	112	256	260	270
	280	299	308	
entrepreneurs	38			
entropy	107	134	139	170
	186	237		
environment	7	17	19	34
	37	68	81	139
	204	205	213	221
	270	332	365	428
	497	502	533	538
	541			
Environmental Defense Fund	204			
environmental justice	18	20	43	
equity	70	75	463	
<i>see also</i> environmental justice				
European Space Agency	416			
European Union	2	20	37	77
	93	138		
Commission	138	285	416	
externalities	238	270	430	662

## Index Terms

## Links

### F

FedEx	24	506		
ferries	211	302		
four-stage traffic models	162	238	609	
Frank–Wolfe algorithm	594			
free trade	2			
freight models	200	211	238	278
	339	352	357	453
	457	462	507	
fuel costs	170	265		

### G

Galileo	412	416	541	
game theory	567			
general equilibrium models	187			
computable	269			
General Motors Corporation	455	496		
generalized costs	192	279	323	
generalized distance	279			
geocoding	376	378		
geographic information systems	3	8	22	50
	51	139	152	159
	222	225	293	299
	300	305	309	329
	357	375	391	411
	434	470	473	493
	550	648	658	
devices	433			
global navigation satellite systems	411			

This page has been reformatted by Knovel to provide easier navigation.

## Index Terms

## Links

Global Positioning System	3	8	325	357
	374	375	386	399
	407	411	433	451
	469	489	506	509
	632	648	655	
Global System for Mobile Communications	5	389	454	
GLONASS	412	416	51	
governance	34			
GPS, <i>see</i> Global Positioning System				
Granger causality	89			
gravity model	14	17	169	186
	238	285	516	551

## **H**

hazardous cargoes	455			
hedonic indices	88			
highways, <i>see</i> roads				
Hong Kong Transport Department	461	619		
housing	44	50	129	259
	262	280	332	
hub-and-spoke systems	28	77		

## **I**

impact analysis	17	83		
information and computer technology	30	278	310	355
	509			
<i>see also</i> geographic information systems				

<u>Index Terms</u>	<u>Links</u>			
infrastructure	28	41	45	50
	77	130	151	200
	237	272	285	365
input–output analysis	100	136	155	187
	198	203	223	
insurance	463			
intelligent transport systems	3	7	24	31
	370	398	451	464
	489			
intelligent highway system	508			
intelligent speed adaptation	453	454	459	465
International Organization for Standardization	370			
interviews, <i>see</i> surveys				
in-vehicle navigation systems	457			
<b>J</b>				
just-in-time systems	103			
<b>K</b>				
Kurth	92			
<b>L</b>				
land use forecasts, <i>see</i> land use modeling				
land use modeling	168	185	205	255
	294	335	355	580
	609			

## Index Terms

## Links

land use planning	7	37	39	68
	115	171	205	255
	330	375	489	544
	629			
land use transportation models	127	147	168	185
	255	334	345	355
	609			
light rail, <i>see</i> transit				
linear programming	223	231		
location-based services	656			
location awareness technology	655			
logit models	134	157	610	613
	620			
linked	526	623		
multinomial	241	250	525	
nested	525			
Lowry models	168	203		
Lowry–Garin models	177			
Lyapunov experiments	249			

## **M**

mapping	414	428	453	457
	463	471	493	498
	535			
<i>see also</i> cognitive maps				
marketing	280			
mathematical programming	568	576	580	
mental maps, <i>see</i> cognitive maps				
migration	13			
mobile phones	458			

This page has been reformatted by Knovel to provide easier navigation.

## Index Terms

## Links

mode choice	187	198	238
monocentric cities	204		
monopoly power	18		
multi-nuclei cities	28		
multiplier effects	83	87	

## **N**

National Marine Electronics Association	417			
National Road Transport Commission	455			
NAVSTAR	411	414		
near-neighbour analysis	367			
network design	23	173		
networks	120	128	153	238
	300	310	314	325
	339	584	653	
artificial neural	247			
design	23	173	300	346
	392	457	502	524
	567			
hyper	281			
micro	303			
neural	238	550		
potential network area	653			
spatial equilibrium	583			
time prism	653	655		
new economic geography	2	8	222	
new growth theory	84	284		
new institutional economics	33			
new transport geography	9	13		
new urban economics	2			

This page has been reformatted by Knovel to provide easier navigation.

## Index Terms

## Links

noise nuisance	139	270
non-linear dynamic systems	238	
North-American Free Trade Area	2	

## **O**

object-oriented programming languages	550
Open Skies policy	89

## **P**

parking	112	119	195	204
	261	554		
peaked demand	196	476	643	
pedestrians	68	119	122	213
	218	261	263	264
	304	433	440	445
	518	528	551	557
	560	639		
ports	17			
potential path trees	651			
predator–prey systems	242	245		
price discrimination	279			
probit models	610	615	621	
public transport, <i>see</i> transit	207	217		
Puget Sound Regional Council				

## **Q**

queueing	555
----------	-----

## Index Terms

## Links

### R

railways	14	68	103	119
	196	217	233	260
	263	283	302	304
	311	316	339	392
	502	524		
random utility models	157	186	196	
real time analysis	3	369	456	
rents	28	134	168	187
		193		
residential density	67	152	169	
revealed preference models	127	142	196	
risk assessment	47			
road haulage	19	196	198	218
		357		
road pricing	69	201	204	213
		461		
roads	13	15	49	55
	68	85	204	211
	213	259	300	305
	307	317	319	320
	339	346	365	376
	392	441	452	503
	508	556	584	
high-occupancy lanes	324	339		
intersections	395			
US Interstate Highway System	2	15	43	

<u>Index Terms</u>	<u>Links</u>			
route assignment	181	187	198	295
	298	350	357	361
	367	370	391	402
	475	554	580	587
	600	609		
dynamic	611			
Russian Ministry of Defense	416			
<b>S</b>				
safety	211	321	454	456
	460	495	507	511
	539	556		
satisficing	186			
security, <i>see</i> terrorism				
scenario analysis	551			
scheduling	474			
school trips	195	215	257	456
	514	520	632	
Service Applications International	92			
shipping	17	68	417	563
shopping	169	195	261	289
	336	364	433	519
	523	543	559	630
	632	644	648	
Sierra Club	204	210		
simulation models	168	203	238	269
	402	524		
microsimulations	256	223	325	353
	554			
Monte Carlo	527	618		

This page has been reformatted by Knovel to provide easier navigation.

## Index Terms

## Links

Singapore Land Transport Authority	461				
social exclusion	139				
spatial clusters	97	364			
spatial equilibrium	583				
speed	215	459	471	556	
	560	579			
sprawl	30	32	34	37	
Standing Advisory Committee on Trunk Road					
Assessment	284				
stated preference models	127	142			
stochastic user equilibrium	610	613			
subsidies	199				
suburbs	30	72	433	524	
surveys	85	127	375	441	
	487				
<i>see also</i> activity analysis and geographic information systems					
sustainability	17	19	33		

## **T**

traveling-salesman problem	391	400	404		
taxation	30	149	161	199	
	212				
telebanking	31				
telecommunications	153	237	243	583	
	599				
telematics	3				
teleshopping	31				
teleworking	24	31	599		
terrorism	34	611			

This page has been reformatted by Knovel to provide easier navigation.

## Index Terms

## Links

Tinbergen–Bos systems	97			
tolls	170	195	259	543
	459	461	571	579
tour-based models	353	513		
tourism	285			
toxic release inventories	45	50		
trade and location	188	285		
traffic assignment, <i>see</i> route assignment				
traffic calming	213			
traffic forecasting	17	167	180	312
	336			
traffic generation	3	609		
traffic signals	571	611		
trains, <i>see</i> railways				
Trans-European Transport Networks	285			
transit	3	68	72	74
	111	149	150	161
	195	217	259	262
	264	265	304	320
	347	384	440	457
	453	455	456	503
	521	527	554	639
transportation planning	2	5	15	294
	337	434	444	479
	559	609	629	632
Transportation Research Board	20	312	319	427
	434	469	477	
travel forecasting	167	203	294	
travel time values	175	177	196	629
trip chains	513	644		

This page has been reformatted by Knovel to provide easier navigation.

## Index Terms

## Links

trip lengths	175	397		
trip timing	113	117		
truckling, <i>see</i> road haulage				
<b>U</b>				
UK Civil Aviation Authority	92			
UPS	24	506		
urban form	1	24	28	157
	168	185	255	331
		334		
URS Corporation	21			
US Bureau of Census	73	311	332	
US Bureau of Labor Statistics	634			
US Bureau of Public Roads	15			
US Bureau of Transportation Statistics	23	313	344	
US Department of Commerce	294			
US Department of Defense	411	412	413	414
	416	421	423	
US Department of Housing and Urban Development	70	332		
US Department of Transportation	30	111	114	118
	123	313	376	413
	426	427	455	
US Environmental Protection Agency	46	429	540	541
US Federal Aviation Administration	80			
US Federal Communications Commission	427			
US Federal Highways Administration	434	489	490	492
	495	496	497	
US Federal Highways Administration	49	68	310	352
US General Accounting Office	80			

This page has been reformatted by Knovel to provide easier navigation.

## Index Terms

## Links

US National Household Travel Survey	257	260
US National Science Foundation	20	
US Office of Technology Assessment	30	

## **V**

vehicle managed inventory	457
vehicle routing, <i>see</i> route assignment	
Virginia Department of Transportation	492

## **W**

walking, <i>see</i> pedestrians				
Wardrop's principles	585	589	610	612
warehousing	358	464	369	656
work trips	169			
World Commission on Environment and Development	19			

## **Z**

zones/zoning	37	39	136	149
	161	172	186	204
	213	216	294	303
	330	335	344	378
	520	551		